11-1-2012

Vol. 11, No. 2 (Full Issue)

JMASM Editors

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

Recommended Citation

Editors, JMASM (2012) "Vol. 11, No. 2 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 11 : Iss. 2 , Article 26.

DOI: 10.22237/jmasm/1351743900

Available at: <http://digitalcommons.wayne.edu/jmasm/vol11/iss2/26>

This Full Issue is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

```

do i1=1,4
  j(1)=i1
  do i2=1,4
    j(2)=i2
    do i3=1,4
      j(3)=i3
      do i4=1,4
        j(4)=i4
        if (j(1) .eq. j(2) .or. j(1) .eq. j(3) .or. j(1) .eq. j(4)) cycle
        if (j(2) .eq. j(3) .or. j(2) .eq. j(4)) cycle
        if (j(3) .eq. j(4)) cycle
        print*,j(1),j(2),j(3),j(4)
      end do
    end do
  end do
end do

```

Journal of Modern Applied Statistical Methods

Invited Article

296 – 302

Rand R. Wilcox

Comparing Two Independent Groups
Via a Quantile Generalization of the
Wilcoxon-Mann-Whitney Test

Shlomo S. Sawilowsky
Evaluation & Research
Wayne State University
Editor

Harvey J. Keselman
Department of Psychology
University of Manitoba
Associate Editor

Bruno D. Zumbo
Meas., Eval., & Res.
Univ. of British Columbia
Associate Editor

Vance W. Berger
Biometry Research Group
National Cancer Institute
Assistant Editor

John L. Cuzzocrea
Educational Research
University of Akron
Assistant Editor

Todd C. Headrick
Educational Psychology
Southern Illinois University
Carbondale
Assistant Editor

Alan Klockars
Educational Psychology
University of Washington
Assistant Editor

Journal Of Modern Applied Statistical Methods

Shlomo S. Sawilowsky

Editor

College of Education
Wayne State University

Harvey Keselman

Associate Editor

Department of Psychology
University of Manitoba

Bruno D. Zumbo

Associate Editor

Measurement, Evaluation, & Research Methodology
University of British Columbia

Vance W. Berger

Assistant Editor

Biometry Research Group
National Cancer Institute

John L. Cuzzocrea

Assistant Editor

Educational Research
University of Akron

Todd C. Headrick

Assistant Editor

Educational Psychology and Special Education
Southern Illinois University-Carbondale

Alan Klockars

Assistant Editor

Educational Psychology
University of Washington

Journal Of Modern Applied Statistical Methods

Invited Article

296 – 302	Rand R. Wilcox	Comparing Two Independent Groups Via a Quantile Generalization of the Wilcoxon-Mann-Whitney Test
-----------	-----------------------	--

Regular Articles

303 – 318	Jay R. Schaffer, Shawn VandenHul	A Graphical Examination of Variable Deletion within the MEWMA Statistic
319 – 324	Stan Lipovetsky	Regression Split by Levels of the Dependent Variable
325 – 335	Shipra Banik, B. M. Golam Kibria, Dinesh Sharma	Testing the Population Coefficient of Variation
336 – 347	Diwakar Shukla, Sharad Pathak, Narendra Singh Thakur	Class(es) of Factor-Type Estimator(s) in Presence of Measurement Error
348 – 360	Guven Ozkaya, Ilker Ercan	Examining Multiple Comparison Procedures According to Error Rate, Power Type and False Discovery Rate
361 – 367	R. Sampath Kumar, M. Indra, R. Radhakrishnan	Selection of Mixed Sampling Plan with QSS-1($n; C_N, C_T$) Plan as Attribute Plan Indexed Through MAPD and LQL
368 – 377	Anwar Hassan, Sheikh Bilal	On Some Negative Integer Moments of Quasi-Negative-Binomial Distribution
378 – 384	Anwar Hassan, Sheikh Bilal, Imtiyaz Ahmad Shah	On Some Properties of Size-Biased Polya-Eggenberger Distribution
385 – 395	S. A. Al-Subh, M. T. Alodat, Kamaruzaman Ibrahim, Abdul Aziz Jemain	Modified EDF Goodness of Fit Tests for Logistic Distribution under SRS and RSS
396 – 406	R. Sankle, J. R. Singh	Single Sampling Plans for Variables Indexed by AQL and AOQL with Measurement Error

407 – 415	Abdullah A. Smadi, Nour H. Abu-Afouna	An Extension of Cochran-Orcutt Procedure for Generalized Linear Regression Models with Periodically Correlated Errors
416 – 442	Adil Hamid Khan, T. R. Jan	Bayesian Estimation of Erlang Distribution under Different Generalized Truncated Distributions as Priors
443 – 449	Ghadban Khalaf	A Proposed Ridge Parameter to Improve the Least Squares Estimator
450 – 456	Shyam S. Ganguly	Exact Logistic Regression for a Matched Pairs Case-Control Design with Polytomous Exposure Variables
457 – 468	Munni Begum, Jay Bagga, C. Ann Blakey	Graphical Modeling for High Dimensional Data
469 – 478	Karim K. Mardaneh	Small-to-Medium Enterprises and Economic Growth: A Comparative Study of Clustering Techniques
479 – 492	Deniz Sigirli, Ilker Ercan	Examining Growth with Statistical Shape Analysis and Comparison of Growth Models
493 – 505	B. Srinivasa Rao, J. Pratapa Reddy, G. Sarath Babu	Extreme Value Charts and Analysis of Means (ANOM) Based on the Log Logistic Distribution
506 – 516	Gaetano Ferrieri	Ferrieri's Index of Openness Applied to Remittances to Developing Countries
517 – 523	Hakim Bekrizadeh, Gholam Ali Parham, Mohamd Reza Zadkarmi	Weighted Cook-Johnson Copula and their Characterizations: Application to Probable Modeling of Hot Spring Eruptions
524 – 529	Adewara Johnson Ademola, Mbata Ugochukwu Ahamefula	Multivariate Generalized Poisson Distribution for Inference on Selected Non-Communicable Diseases in Lagos State, Nigeria
<i>Brief Report</i> 530 – 535	Raja Sultan, S. P. Ahmad	Posterior Estimates of Poisson Distribution Using R Software

JMASM is an independent print and electronic journal (<http://www.jmasm.com/>), publishing (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Editorial Assistant: **Julie M. Smith, Ph. D.**

Invited Article
**Comparing Two Independent Groups
Via a Quantile Generalization of the Wilcoxon-Mann-Whitney Test**

Rand R. Wilcox
University of Southern California,
Los Angeles, CA USA

The Wilcoxon-Mann-Whitney test, as well as modern improvements, are based in part on an estimate of $p = P(D < 0)$, where $D = X - Y$ and X and Y are independent random variables; a common goal is to test $H_0: p = 0.5$. This corresponds to testing $H_0: \xi_{0.5}$, where $\xi_{0.5}$ is the 0.5 quantile of the distribution of D . If the distributions associated with X and Y do not differ, then D has a symmetric distribution about zero. In particular, $\xi_q + \xi_{1-q} = 0$ for any $q \leq 0.5$, where ξ_q is the q^{th} quantile. Methods aimed at testing $H_0: p = 0.5$ are generalized by suggesting a method for testing $H_0: \xi_q + \xi_{1-q} = 0, q < 0.5$

Key words: Bootstrap methods, Harrell-Davis estimator, tests for symmetry, tied values, Well Elderly study.

Introduction

Consider two independent random variables, X and Y , let $D = X - Y$ and let τ_d , τ_x and τ_y be the population medians of D , X and Y , respectively. It is known that, under general conditions, the Wilcoxon-Mann-Whitney (WMW) test does not test $H_0: \tau_x = \tau_y$ (Fung, 1980). The WMW test is based on an estimate of $p = P(X < Y)$, but under general conditions it uses the wrong standard error, in contrast to more modern methods aimed

at correcting this problem (Cliff, 1996; Brunner Munzel, 2000; Newcombe, 2006a, 2006b). The explicit goal of these improvements is making inferences about p , which includes the common goal of testing

$$H_0: p = 0.5. \quad (1)$$

Moreover, it is known, and fairly evident, that testing (1) corresponds to testing

$$H_0: \tau_d = 0. \quad (2)$$

Rand R. Wilcox is a Professor of Psychology. He is the author of eight textbooks on statistics, including *Modern Statistics for the Social and Behavioral Sciences* (2012, New York, CRC Press). Email him at: rwilcox@usc.edu.

Inferences about p and τ_d are important and useful, but a deeper understanding of how two independent groups compare would result by knowing something about the quantiles of the distribution of D .

For illustrative purposes, imagine that some experimental method is being compared to a control group and that $D > 0$ indicates that the experimental method is more effective than no treatment. If D has a skewed distribution, it is possible that p is approximately 0.5 and that testing (1) has relatively low power, yet there is a sense in which the experimental method is beneficial. Let ξ_q be the q^{th} quantile of D and assume, for example, that $\xi_{0.25} = -4$ and $\xi_{0.75} = 6$. Thus, for randomly sampled observations from each group, there is a sense in which the experimental treatment outweighs no treatment. If there are no benefits, then D should have a symmetric distribution about zero. In particular, it should be the case that

$$H_0: \xi_q + \xi_{1-q} = 0 \quad (3)$$

is true for any $q \leq 0.5$; consequently, this article suggests a method for testing (3).

Note that information about $\xi_q + \xi_{1-q}$ for a range of q values provides a more detailed sense about the distribution of D compared to using a single measure of location. For example, a portion of the study conducted by Jackson, et al. (2009) dealt with assessing the extent a particular intervention strategy reduced depression in older adults. An issue is whether the efficacy of the intervention changes as an individual moves from the center of the distribution of D to the tails. For the Jackson, et al. (2009) study, an estimate of the 0.9 quantile is 27.6 and the estimate of the 0.1 quantile is -19.7. That is, the drop in depression, 27.6, as reflected by the 0.9 quantile, exceeds the increase in depression, as reflected by the estimate of the 0.1 quantile, -19.7. For the 0.4 and 0.6 quantiles, the estimates are -1 and 5, again suggesting that intervention is useful, but the impact of intervention is less striking. If the distributions differ in terms of a measure of location only, it would be the case that $\xi_q + \xi_{1-q}$ does not vary with q .

For completeness, Wilcox and Erceg-Hurn (in press) considered the case where X and Y are dependent with two goals. The first is to compare the quantiles of the marginal distributions and the other is to test (3) but with D corresponding to the usual paired differences.

Note that this differs from the situation at hand. For dependent groups, the goal is to assess changes within a subject in terms of the quantiles of D ; here, the goal is make inferences about the difference between two randomly sample participants. A crude description of the method by Wilcox and Erceg-Hurn is that it generalizes the sign test for dependent groups. The suggestion is that a similar generalization of the Wilcoxon-Mann-Whitney test might be of interest. (Note that control over the Type I error probability is a function of both q and the sample sizes.) It was found that conditions under which good control over the Type I error probability is achieved differ to some degree from those when comparing dependent groups.

Description of the Proposed Method

A variety of methods for estimating the q^{th} quantile have been proposed, comparisons of which are reported by Parrish (1990), Sheather and Marron (1990) and Dielman, Lowry and Pfaffenberger (1994). The simplest approach is to estimate the q^{th} quantile using a single order statistic. Another approach is to use an estimator based on a weighted average of two order statistics while other estimators are based on a weighted average of all the order statistics. Regarding the issue of which estimator is best, the only certainty is that no single estimator dominates in terms of efficiency. For example, the Harrell and Davis (1982) estimator has a smaller standard error than the usual median when sampling from a normal distribution or a distribution that has relatively light tails, but for sufficiently heavy-tailed distributions, the reverse is true (Wilcox, 2012, p. 87).

Consider the special case where the goal is to estimate the population median. Currently all methods that are based in part on an estimate of the standard error of the usual sample median can perform poorly when tied values occur (Wilcox, 2006).

There are two problems: The first is obtaining a reasonably accurate estimate of the standard error. Many estimators have been proposed, all of which can be highly inaccurate when there are tied values. The second general concern is that, when tied, values occur the usual sample median is not necessarily asymptotically normal. Wilcox (2012) illustrated this result

when the cardinality of a sample space is relatively small. To date, the only method known to perform reasonably well in simulations is a slight generalization of the standard percentile bootstrap method (Wilcox, 2006). Thus, an obvious speculation is that when the goal is to make inferences about the quantiles of the distribution associated with D , the same percentile bootstrap method might perform well. However, simulations indicate that this is not necessarily the case.

Let n_j be the sample size for the j^{th} group ($j = 1, 2$). Consider, for example, the situation where $n_1 = 20$, $n_2 = 30$ and observations are generated from a binomial distribution with probability of success 0.4 and when the sample space is $0(1)7$. When testing at the 0.05 level, simulations indicate that the actual level is approximately 0.102. Due to the difficulty of not being able to get a reasonably accurate estimate of the standard error when sampling from a discrete distribution, bootstrap methods based in part on an estimate of the standard error hold little promise.

Here, the one method that performed well in simulations was based in part on the estimator derived by Harrell and Davis (1982) that estimates the q^{th} quantile using a weighted average of all the order statistics. More precisely, let Y be a random variable having a beta distribution with parameters $a = (n + 1)q$ and $b = (n + 1)(1 - q)$. That is, the probability density function of Y is

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1} (1-y)^{b-1},$$

where Γ is the gamma function.

Let

$$W_i = P((i-1)/n \leq Y \leq i/n).$$

For the random sample $X_1, \dots, X_n$, let $X_{(1)} \leq \dots \leq X_{(n)}$ denote the observations written in ascending order. The Harrell-Davis estimate of ξ_q is $\hat{\xi}_q = \sum W_i X_{(i)}$. In terms of its standard error, Sfakianakis and Verginis (2006) show that in some situations the Harrell-Davis estimator competes well with alternative estimators that

use a weighted average of all the order statistics, but there are exceptions. For example, Sfakianakis and Verginis (2006) derived alternative estimators that have advantages over the Harrell-Davis in some situations, but it was found that when sampling from heavy-tailed distributions the standard errors of their estimators can be substantially larger than the standard error of Harrell-Davis estimator.

To describe the details of the proposed test of (3), let $X_1, \dots$, and $Y_1, \dots$, be random samples of size n_1 and n_2 , respectively, and let $D_{ik} = X_i - Y_k$ ($i = 1, \dots, n_1$; $k = 1, \dots, n_2$). The q^{th} quantile of distribution of D , δ_q , is estimated via the Harrell-Davis estimator, applied to the D_{ik} values, yielding $\hat{D}_q$. Next, generate a bootstrap sample from the j^{th} group by resampling with replacement n_j observations from group j . Let $\tilde{D}_q$ be the estimate of q^{th} quantile of D based on these bootstrap samples and let $d = \tilde{D}_q + \tilde{D}_{1-q}$. Repeat this process B times yielding d_b , $b = 1, \dots, B$; here, $B = 1,000$ is used. Let $\ell = \alpha B/2$, rounded to the nearest integer, and let $u = B - \ell$. Letting $d_{(1)} \leq \dots \leq d_{(B)}$ represent the B bootstrap estimates written in ascending order, an approximate $1 - \alpha$ confidence interval for $\delta_q + \delta_{1-q}$ is $(d_{(\ell+1)}, d_{(u)})$. This will be called method DHD.

Let A denote the number of times d is less than zero and let C be the number of times $d = 0$. Letting

$$\hat{p} = \frac{A + .5C}{B},$$

a (generalized) p -value is $2\min(\hat{p}, 1-\hat{p})$ (Liu & Singh, 1997).

Results

Simulations were used to study the small-sample properties of method DHD. The sample sizes considered were $(n_1, n_2) = (10, 10), (20, 20), (10, 30)$ and $(20, 30)$. Estimated Type I error probabilities were based on 2,000 replications. Two values for q were considered: 0.25 and 0.1. Both continuous and discrete distributions were used. The four continuous distributions were normal, symmetric and heavy-tailed, asymmetric and light-tailed and asymmetric and heavy-tailed. More precisely, four g -and- h distributions

were used (Hoaglin, 1985) that contain the standard normal distribution as a special case. If Z has a standard normal distribution, then

$$W = \frac{\exp(gZ) - 1}{g} \exp\left(h \frac{Z^2}{2}\right), \text{ if } g > 0,$$

$$= Z \exp\left(h \frac{Z^2}{2}\right), \text{ if } g = 0$$

has a g -and- h distribution where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0$, $g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 1 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Additional properties of the g -and- h distribution are summarized by Hoaglin (1985).

Table 1: Some Properties of the g -and- h Distribution

g	h	κ_1	κ_2
0.0	0.0	0.0	3.0
0.0	0.2	0.0	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

To gain perspective on the effects of tied values, data were generated from a discrete distribution having a sample space consisting of the integers 0 through 7; more precisely, data were generated from a binomial distribution with probability of success equal to 0.4. First consider the four g -and- h distributions when testing at the 0.05 level and $n_1 = n_2 = 10$. As indicated in Table 2, if $q = 0.25$, in which case the goal is to test (3) with $q = 0.25$, then $\hat{\alpha}$, the probability of a Type I error, is estimated to be

close to the nominal level. Note that the estimates barely change among the continuous distributions considered. However, when $q = 0.1$, the estimated Type I error probability can exceed 0.1. Increasing one of the sample sizes to 30 improves the estimate, but it still exceeds 0.075. Although the seriousness of a Type I error can depend on the situation, Bradley (1978) suggested that, as a general guide, when testing at the 0.05 level the actual level should not exceed 0.075. With $n_1 = 20$ and $n_2 = 40$, again the estimate can exceed 0.1. With $n_1 = n_2 = 30$ (not shown in Table 2), reasonably accurate control over the probability of Type I error is achieved. Increasing both sample sizes to 40, the probability of Type I error is estimated to be between 0.045 and 0.051 among all situations considered.

Generating data from the binomial distribution gave results similar to those in Table 2. For $n_1 = n_2 = 10$ and $q = 0.25$, $\hat{\alpha} = 0.065$. For $n_1 = n_2 = 20$ $\hat{\alpha} = 0.056$ and 0.063 for $q = 0.25$ and 0.1, respectively. For $n_1 = 20$ and $n_2 = 30$ the estimates are 0.056 for both $q = 0.25$ and $q = 0.1$.

How the power of method DHD compares to other methods depends in part on the nature of the distributions being compared. As is evident, different methods are sensitive to different features of the data. However, to provide at least some perspective, some results are reported when distributions differ in location only. In particular, consider $D = X - Y + \lambda$ for some constant λ where both X and Y have mean zero and variance one. Under normality, it can be seen that $\delta_q + \delta_{1-q} = 2\lambda$. Thus, when comparing means, rather than testing (3), this suggests that method DHD might have relatively high power under normality despite the sample mean having a smaller standard error than the Harrell-Davis estimator. Table 3 reports some simulation power estimates when $q = 0.25$. The column headed by Welch indicates the estimated power when using the method from Welch (1938) to test the hypothesis of equal means. As shown, the power of method DHD compares well to Welch's method – and that DHD seems to have a slight advantage.

A QUANTILE GENERALIZATION OF THE WILCOXON-MANN-WHITNEY TEST

Table 2: Estimated Type I Error Probability, $\alpha = 0.05$

q	n_1	n_2	g	h	$\hat{\alpha}$
0.25	10	10	0.0	0.0	0.069
			0.0	0.2	0.066
			0.2	0.0	0.072
			0.2	0.2	0.073
	20	20	0.0	0.0	0.060
			0.0	0.2	0.056
			0.2	0.0	0.060
			0.2	0.2	0.058
	10	30	0.0	0.0	0.082
	20	30	0.0	0.0	0.062
0.10	10	10	0.0	0.0	0.092
			0.0	0.2	0.104
			0.2	0.0	0.091
			0.2	0.2	0.108
	20	20	0.0	0.0	0.065
			0.0	0.2	0.069
			0.2	0.0	0.065
			0.2	0.2	0.067
	20	30	0.0	0.0	0.060

An Illustration

Consider the Jackson, et al. (2009) study described in the introduction that used sample sizes of 232 and 140. Figure 1 shows an estimate of $\delta_q + \delta_{1-q}$, indicated by *, as a function of q, where the q values are 0.05(0.05)0.40. The corresponding p-values are 0.002, 0.004, 0.008, 0.010, 0.016, 0.020, 0.020 and 0.020. The + above and below the * indicate a 0.95 confidence interval. These results suggest that intervention is effective and that this is the case particularly in terms of more extreme quantiles.

Conclusion

In terms of controlling the probability of a Type I error, method DHD generally performs well in simulations. The restriction is that as q approaches zero larger samples size are needed, particularly when the sample sizes are unequal.

For $n_1 = n_2 = 10$, all indications are that method DHD performs reasonably well for $q \geq 0.2$. For $n_1 = 10$ and $n_2 = 30$, this is not the case, however, for $\min(n_1, n_2) \geq 20$, control over the Type I error probability was found to be reasonably satisfactory.

It is not suggested that method DHD should be used to the exclusion of all other techniques aimed at comparing two independent groups. Rather, the suggestion is that multiple techniques are needed to obtain a good understanding of how two groups compare and the DHD method helps achieve this goal.

Finally, method DHD can be applied with the R function `cbmhd`. The R function `qwmwd` applies the method using a range of q values. The plot in Figure 1 was created with the latter function.

RAND R. WILCOX

Table 3: Estimated Power, $\alpha = 0.05$, $\lambda = 1$

q	n_1	n_2	g	h	DHD	WELCH
0.25	10	10	0.0	0.0	0.62	0.55
			0.0	0.2	0.60	0.54
			0.2	0.0	0.43	0.36
			0.2	0.2	0.42	0.35

Figure 1: Estimates of $\xi_q + \xi_{1-q}$

These R functions are included in a package that can be downloaded from <http://college.usc.edu/labs/rwilcox/home>.

References

- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Brunner, E. & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and small-sample approximation. *Biometrical Journal*, 42, 17-25.
- Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Erlbaum.
- Dielman, T., Lowry, C. & Pfaffenberger, R. (1994). A comparison of quantile estimators. *Communications in Statistics-Simulation and Computation*, 23, 355-371.
- Fung, K. Y. (1980). Small sample behaviour of some nonparametric multi-sample location tests in the presence of dispersion differences. *Statistica Neerlandica*, 34, 189-196.
- Harrell, F. E. & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, 69, 635-640.
- Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In *Exploring Data Tables Trends and Shapes*, D. Hoaglin, F. Mosteller & J. Tukey (Eds.), 461-515. New York, NY: Wiley.
- Jackson, J., Mandel, D., Blanchard, J., Carlson, M., Cherry, B., Azen, S., Chou, C.-P., Jordan-Marsh, M., Forman, T., White, B., Granger, D., Knight, B., & Clark, F. (2009). Confronting challenges in intervention research with ethnically diverse older adults: The USC Well Elderly II trial. *Clinical Trials*, 6, 90-101.
- Liu, R. G. & Singh, K. (1997). Notions of limiting P values based on datadepth and bootstrap. *Journal of the American Statistical Association*, 92, 266-277.
- Newcombe, R. G. (2006a). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: General issues and tail-area-based methods. *Statistics in Medicine*, 25, 543-557.
- Newcombe, R. G. (2006b). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: Asymptotic methods and evaluation. *Statistics in Medicine*, 25, 559-573.
- Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics*, 46, 247-257.
- Sfakianakis, M. E. & Verginis, D. G. (2006). A New Family of Nonparametric Quantile Estimators. *Communications in Statistics-Simulation and Computation*, 37, 337-345.
- Sheather, S. J. & Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, 85, 410-416.
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350-362.
- Wilcox, R. R. (2006). Comparing medians. *Computational Statistics & Data Analysis*, 51, 1934-1943.
- Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing (3rd Edition)*. San Diego, CA: Academic Press.

Regular Articles

A Graphical Examination of Variable Deletion within the MEWMA Statistic

Jay R. Schaffer Shawn VandenHul
University of Northern Colorado
Greeley, CO

A general procedure for identifying the variable(s) that contribute(s) to the signal of the multivariate extension of the exponentially weighted moving average (MEWMA) chart is presented. The procedure systematically removes one or two variables from the MEWMA statistic calculations. Percentages are calculated for correctly identifying various shifts.

Key words: Multivariate quality control, MEWMA, variable deletion.

Introduction

With modern computers, it is common to monitor several correlated quality characteristics simultaneously. Various types of multivariate control charts have been proposed to take advantage of the relationships among variables being monitored (Alt, 1984; Jackson, 1985; Wierda, 1994; Lowry and Montgomery, 1995; Mason, et al., 1997). Lowry, et al. (1992) proposed a multivariate extension of the exponentially weighted moving average (MEWMA) control chart. They demonstrated that the average run length (ARL) performance of the MEWMA is similar to that of the multivariate cumulative sum (MCUSUM) control charts discussed by Crosier (1988) and Pignatiello and Runger (1990) and is better than Hotelling's (1947) χ^2 chart for detecting a shift in the mean vector of a multivariate normal distribution.

Woodall and Montgomery (1999) showed that, once an out-of-control signal is given by a multivariate chart, it may be difficult to identify the variable (or variables) that

contributed to the signal. Jackson (1980, 1991) proposed examining the Hotelling's T^2 statistic (Jackson, 1985) using principle component analysis (PCA). Mason, et al. (1995) suggested decomposing Hotelling's T^2 statistic by removing individual variables from its calculation. Woodall and Montgomery (1999) noted that additional work is needed on graphical methods for data visualization when interpreting signals from multivariate control charts.

This article presents a graphical approach to identify the source of a signal from the MEWMA control chart and examines the effects of systematically deleting a variable, or pairs of variables, from the calculations of the MEWMA statistic. The methodology is similar to examining the PRESS residuals (Allen, 1971) or DFBETAS (Belsley, et al., 1980) in regression analysis. Methodology used herein deletes variables in a multivariate process as opposed to deleting individual observations in a data set; in addition, the probability of correctly identifying the source using various simulations is estimated.

MEWMA Chart

Assume a sequence of independent observations from a p -variate normal distribution whose mean vector shifts from μ_0 to μ_1 on the r^{th} observation, that is,

Jay Schaffer is an Associate Professor of Statistics and is currently the chair of Applied Statistics and Research Methods. Email him at: jay.schaffer@unco.edu. Shawn VandenHul is a graduate of the Ph.D. program in Applied Statistics and Research Methods.

GRAPHICAL EXAMINATION OF VARIABLE DELETION WITHIN THE MEWMA

$$\begin{aligned} \mathbf{x}_i &\sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}), \quad i = 1, 2, \dots, r-1 \\ &\sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad i = r, r+1, r+2, \dots \end{aligned} \quad (1)$$

Lowry, et al. (1992) defined vectors of exponentially weighted moving averages,

$$\mathbf{z}_i = \lambda \mathbf{x}_i + (1 - \lambda) \mathbf{z}_{i-1} \quad (2)$$

$i = 1, 2, \dots$, where $\mathbf{z}_0 = \mathbf{0}$ and $0 < \lambda \leq 1$. The MEWMA chart would give an out-of-control signal if

$$T_i^2 = \mathbf{z}_i' \boldsymbol{\Sigma}_{\mathbf{z}_i}^{-1} \mathbf{z}_i > h \quad (3)$$

where $h > 0$ is chosen to achieve a specified in-control ARL and

$$\boldsymbol{\Sigma}_{\mathbf{z}_i} = \frac{\lambda}{2 - \lambda} \left[1 - (1 - \lambda)^{2i} \right] \boldsymbol{\Sigma}. \quad (4)$$

The one-variable deletion within the MEWMA statistic removes variables from the T_i^2 statistic when a signal is detected. This study examines the removal of one variable and two variables at a time: one-variable deletion removes one variable at a time and recalculates the current T_i^2 statistic excluding the removed variable, two-variable deletion removes pairs of variables and recalculates the current T_i^2 statistic excluding the removed pair. Given either method, a small, reduced T_i^2 statistic would indicate a possible signal source.

One-Variable Deletion

Assume on the s^{th} sample, the MEWMA chart signaled a change ($T_s^2 > h$). The p variables are removed, one at a time, from the calculation of T_s^2 . Assume the j^{th} variable is removed such that

$$\begin{aligned} \mathbf{x}_i' &= (x_{i1}, x_{i2}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{ip}, x_{ij}) \\ &= (\mathbf{x}_{i(j)}', x_{ij}) \end{aligned}$$

where $\mathbf{x}_{i(j)}'$ is a $(p-1) \times 1$ vector excluding the j^{th} variable. In addition, let $\boldsymbol{\Sigma}_{(j)}$ be the $(p-1) \times (p-1)$ principal sub-matrix of $\boldsymbol{\Sigma}$ excluding the j^{th} variable. With the j^{th} variable removed, the MEWMA equations become

$$\mathbf{z}_{i(j)} = \lambda \mathbf{x}_{i(j)} + (1 - \lambda) \mathbf{z}_{i-1, (j)} \quad (5)$$

$i = 1, 2, \dots, s$ where $\mathbf{z}_{0(j)} = \mathbf{0}$,

$$T_{i(j)}^2 = \mathbf{z}_{i(j)}' \boldsymbol{\Sigma}_{\mathbf{z}_{i(j)}}^{-1} \mathbf{z}_{i(j)} \quad (6)$$

and

$$\boldsymbol{\Sigma}_{\mathbf{z}_{i(j)}} = \frac{\lambda}{2 - \lambda} \left[1 - (1 - \lambda)^{2i} \right] \boldsymbol{\Sigma}_{(j)}. \quad (7)$$

The calculation of $T_{i(j)}^2$ continues until the s^{th} sample.

A graphical comparison of the set of reduced MEWMA statistics $\{T_{s(1)}^2, T_{s(2)}^2, \dots, T_{s(p)}^2\}$ to T_s^2 should aid in identifying the cause of the signal. The smallest reduced MEWMA statistic may indicate which variable contributed to the signal. For example, if the 1st variable shifts, the reduced MEWMA statistics may resemble Figure 1. A similar analysis is required if more than two variables change. For example, the reduced MEWMA statistics may resemble Figure 2, if the 1st and the 2nd variables shift or may resemble Figure 3, if the 1st, 2nd and 3rd variables shift.

Consider a modified example from Lowry, et al. (1992). Assume

$$\begin{aligned} \mathbf{x}_i &\sim N_3(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}), \quad i = 1, 2, \dots, 15 \\ &\sim N_3(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad i = 16, 17, 18, \dots \end{aligned} \quad (8)$$

such that:

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

Figure 1: A General Representation of the Reduced MEWMA Statistics if Variable 1 Shifted

Figure 2: A General Representation of the Reduced MEWMA Statistics if Variables 1 and 2 Shifted

Figure 3: A General Representation of the Reduced MEWMA Statistics if Variables 1, 2 and 3 Shifted

$$\boldsymbol{\mu}_1 = \begin{bmatrix} \sqrt{2} \\ 0 \\ 0 \end{bmatrix},$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

Note that a shift of

$$\delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)} = 3$$

occurred on the 16th sample. Table 1 displays a data simulation of these conditions along with the corresponding MEWMA statistics, T_i^2 . Using $\lambda = 0.10$, and $h = 10.97$ (in-control ARL = 200), the MEWMA chart signaled on the 21st observation such that $T_{21}^2 = 11.3551$. However, it is not apparent which variable changed through an examination of the data or the MEWMA chart.

Using the data from Table 1, the first variable is removed from the calculation of the MEWMA statistic. Variables 2 and 3 are used to recalculate a reduced MEWMA statistic, $T_{i(1)}^2$.

The reduced covariance matrix is then:

$$\boldsymbol{\Sigma}_{(1)} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

The reduced MEWMA statistic, $T_{i(1)}^2$, is calculated for $i = 1, 2, \dots, 21$ and displayed in Table 2. Note that, on the 21st sample, $T_{21(1)}^2 = 0.9358$ represents the reduced MEWMA statistic with the contribution of the first variable removed.

Repeating the one-variable deletion procedure for the remaining two variables, the reduced MEWMA statistic excluding variable 2 is $T_{21(2)}^2 = 11.3282$ and the reduced MEWMA

statistic excluding variable 3 is $T_{21(3)}^2 = 9.0015$. Comparing the three reduced MEWMA statistics to the MEWMA statistic $T_{21}^2 = 11.3551$, it is likely variable 1 contributed to the signal. Figure 4 displays the MEWMA statistic along with the three reduced MEWMA statistics.

Two-Variable Deletion

Assume on the s^{th} sample, the MEWMA chart signaled a change ($T_s^2 > h$). The p variables are removed, two at a time, from the calculation of T_s^2 . Assume the j^{th} and k^{th} variables are to be removed. Now let

$$\begin{aligned} \mathbf{x}_i' &= \begin{pmatrix} x_{i1}, x_{i2}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, \\ x_{i,k-1}, x_{i,k+1}, \dots, x_{ip}, x_{ij}, x_{ik} \end{pmatrix}, \\ &= (\mathbf{x}_{i(j,k)}', x_{ij}, x_{ik}) \end{aligned}$$

where $\mathbf{x}_{i(j,k)}'$ is a $(p-2) \times 1$ vector excluding the j^{th} and k^{th} variables. In addition, let $\boldsymbol{\Sigma}_{(j,k)}$ be the $(p-2) \times (p-2)$ principal sub-matrix of $\boldsymbol{\Sigma}$ excluding the j^{th} and k^{th} variables. With the j^{th} and k^{th} variables removed, the MEWMA equations become

$$\mathbf{z}_{i(j,k)} = \lambda \mathbf{x}_{i(j,k)} + (1 - \lambda) \mathbf{z}_{i-1(j,k)} \quad (9)$$

$i = 1, 2, \dots, s$ where $\mathbf{z}_{0(j,k)} = \mathbf{0}$,

$$T_{i(j,k)}^2 = \mathbf{z}_{i(j,k)}' \boldsymbol{\Sigma}_{\mathbf{z}_{i(j,k)}}^{-1} \mathbf{z}_{i(j,k)} \quad (10)$$

and

$$\boldsymbol{\Sigma}_{\mathbf{z}_{i(j,k)}} = \frac{\lambda}{2 - \lambda} [1 - (1 - \lambda)^{2i}] \boldsymbol{\Sigma}_{(j,k)}. \quad (11)$$

The calculation of $T_{i(j,k)}^2$ is continued until the s^{th} sample.

Table 1: Simulated Process with Corresponding MEWMA Statistics, T_i^2

i	x_1	x_2	x_3	T_i^2
1	0.1307	0.5629	-0.7255	0.7203
2	1.5662	-0.3972	0.5767	2.3382
3	0.5733	1.4400	1.4343	1.9161
4	-0.0342	-0.0966	0.8100	1.6907
5	0.2922	0.0853	-0.3257	1.2270
6	-0.2988	-0.7700	0.3948	1.1400
7	0.1389	0.4851	0.1806	0.8966
8	-0.0184	-0.5328	0.4871	1.3231
9	0.6751	-0.3919	-1.4367	1.1445
10	-2.5591	-1.4792	-2.3697	1.0214
11	-1.8930	0.4438	-0.9319	2.2448
12	-0.4950	0.4710	-0.0471	2.6071
13	-1.1572	0.8478	-0.5695	5.2338
14	0.2098	-0.8472	0.1777	2.7816
15	0.0101	0.1780	0.9616	2.0170
16	1.1233	-0.6925	-1.2685	1.1097
17	0.8364	-1.5027	-0.1821	1.6985
18	0.6587	1.0085	0.5520	0.8009
19	2.3631	2.1432	0.9458	2.0496
20	2.4894	0.2182	-0.2358	6.7361
21	2.3260	0.7702	0.5218	11.3551

Table 2: Reduced MEWMA Statistics, $T_{i(1)}^2$

i	1	2	3	4	5	6	7
$T_{i(1)}^2$	1.6690	0.0192	1.1522	1.4192	0.7580	0.9119	0.7640
i	8	9	10	11	12	13	14
$T_{i(1)}^2$	1.2226	0.0973	0.9894	1.5824	1.4426	2.5004	1.2420
i	15	16	17	18	19	20	21
$T_{i(1)}^2$	0.2939	1.0651	1.3763	0.4595	0.5040	0.6895	0.9358

Figure 4: MEWMA T^2 Statistic and the Three Reduced MEWMA Statistics

A graphical comparison of the set of reduced MEWMA statistics $\{T_{s(1,2)}^2, T_{s(1,3)}^2, \dots, T_{s(1,p)}^2, T_{s(2,3)}^2, T_{s(2,4)}^2, \dots, T_{s(p-1,p)}^2\}$ to T_s^2 should aid in identifying the cause of the signal. The smallest group of reduced MEWMA statistics may indicate which variable contributed to the signal. For example, if the 1st variable shifts, the reduced MEWMA statistics may resemble Figure 5, such that the group of reduced MEWMA statistics associated with the first variable is uniformly smaller than the others.

A more detailed analysis is required if two variables have shifted. The smallest reduced MEWMA statistic may indicate which pair of variables changed. In addition, any reduced MEWMA statistic associated with one of the pair of variables that shifted may be slightly larger, yet smaller than any other reduced MEWMA statistic not associated with the pair that changed. The reduced MEWMA statistics may resemble Figure 6, if the 1st and the 2nd variables shift. A similar analysis is required if three variables have shifted. The reduced MEWMA statistics may resemble Figure 7, if the 1st, 2nd and 3rd variables shift.

Consider a modified example from Lowry, et al. (1992). Assume

$$\begin{aligned} \mathbf{x}_i &\sim N_4(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}), \quad i = 1, 2, \dots, 15 \\ &\sim N_4(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad i = 16, 17, 18, \dots \end{aligned} \quad (12)$$

such that:

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} \sqrt{3/2} \\ \sqrt{3/2} \\ 0 \\ 0 \end{bmatrix},$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{bmatrix}.$$

Figure 5: A General Representation of the Reduced MEWMA Statistics if Variable 1 Shifted

Figure 6: A General Representation of the Reduced MEWMA Statistics if Variables 1 and 2 Shifted

Figure 7: A General Representation of the Reduced MEWMA Statistics if Variables 1, 2 and 3 Shifted

GRAPHICAL EXAMINATION OF VARIABLE DELETION WITHIN THE MEWMA

Table 3: Simulated Process with Corresponding MEWMA Statistics, T_i^2

i	X_1	X_2	X_3	X_4	T_i^2
1	0.502	0.130	0.150	0.086	0.296
2	-0.862	-0.877	-0.515	0.025	0.531
3	0.630	1.914	1.396	2.179	2.605
4	0.448	-0.422	-0.036	0.692	2.816
5	-0.995	-0.605	-0.772	-1.700	0.402
6	-0.090	1.305	-1.037	-0.812	1.540
7	-0.951	-1.808	-0.142	0.197	0.772
8	-0.549	-0.136	-0.350	0.671	1.857
9	0.068	-0.312	-2.316	0.680	5.749
10	2.132	0.072	-1.062	-1.362	6.477
11	-0.738	0.141	0.030	1.026	5.355
12	1.293	-1.380	-0.687	0.953	9.455
13	-0.249	-0.954	-1.079	-0.001	11.282
14	0.733	-1.432	0.480	-0.406	10.842
15	0.704	-0.170	-0.120	0.159	10.970
16	2.036	2.011	1.985	1.179	7.910
17	0.950	3.354	1.741	1.824	5.742
18	2.044	0.054	0.281	-0.287	8.560
19	0.858	0.593	0.151	-0.932	9.069
20	1.231	2.576	-0.213	-0.428	13.793

Note that a shift of

$$\delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)} = 3$$

occurred on the 16th sample. Table 3 displays a data simulation of these conditions along with the corresponding MEWMA T_i^2 statistics. Using $\lambda = 0.10$ and $h = 12.93$ (in-control ARL = 200), the MEWMA chart signaled on the 20th observation such that $T_{20}^2 = 13.793$.

Using equations (9)-(11), the reduced MEWMA statistics are $T_{20(1,2)}^2 = 0.296$, $T_{20(1,3)}^2 = 4.771$, $T_{20(1,4)}^2 = 5.213$, $T_{20(2,3)}^2 = 10.481$, $T_{20(2,4)}^2 = 11.246$, and $T_{20(3,4)}^2 = 9.674$. Figure 8 displays the reduced MEWMA statistics. Note that $T_{20(1,2)}^2 = 0.296$ indicates variables 1 and 2 likely contributed to the signal.

Figure 8: MEWMA T^2 Statistic and the Reduced MEWMA Statistics

Methodology

Simulations were conducted to estimate the probability of correctly identifying the source of the MEWMA chart's signal. Consider a sequence of independent observations from a p -variate normal distribution whose mean vector shifts from $\mu_0 = \mathbf{0}$ to μ_1 on the 16th observation, that is,

$$\begin{aligned} \mathbf{x}_i &\sim N_p(\mathbf{0}, \Sigma), \quad i = 1, 2, \dots, 15 \\ &\sim N_p(\mu_1, \Sigma), \quad i = 16, 17, 18, \dots \end{aligned} \quad (13)$$

where

$$\Sigma = \begin{bmatrix} 1 & 0.5 & 0.5 & \dots & 0.5 \\ 0.5 & 1 & 0.5 & \dots & 0.5 \\ 0.5 & 0.5 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1 & 0.5 \\ 0.5 & 0.5 & \dots & 0.5 & 1 \end{bmatrix}_{p \times p}.$$

Forty conditions were examined using $p = 3, 4, 5$ and 10 ; five different μ_1 such that $\delta = 1$; and five different μ_1 such that $\delta = 3$. The vectors μ_1 are constructed such that (1) one variable shifts, (2) two variables shift equally, (3) two variables shift unequally, (4) three variables shift equally or (5) three variables shift unequally. Tables 4 and 5 display the conditions examined such that $\delta = 1$ and $\delta = 3$ respectively. When $p = 10$, approximate decimal values were used in place of exact fractions.

Results

One-Variable Deletion Analysis

If one variable shifts, a one-variable deletion is considered to be a success if the smallest reduced statistic correctly identified the variable that changed. If two variables shift, the one-variable deletion is considered to be a success if the two smallest reduced statistics correctly identify the two variables that changed. If three variables shift, the one-variable deletion is considered to be a success if the three smallest reduced statistics correctly identify the three

GRAPHICAL EXAMINATION OF VARIABLE DELETION WITHIN THE MEWMA

Table 4: Twenty Conditions Examined when $\delta = 1$

p	1 Variable Shift	2 Variable Shift (Equal)	2 Variable Shift (Unequal)	3 Variable Shift (Equal)	3 Variable Shift (Unequal)
3	$\mu_1 = \begin{bmatrix} \sqrt{2/3} \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{1/2} \\ \sqrt{1/2} \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 2\sqrt{2/11} \\ \sqrt{2/11} \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{2/3} \\ \sqrt{2/3} \\ \sqrt{2/3} \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 3\sqrt{1/10} \\ 2\sqrt{1/10} \\ \sqrt{1/10} \end{bmatrix}$
4	$\mu_1 = \begin{bmatrix} \sqrt{5/8} \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{5/12} \\ \sqrt{5/12} \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 2\sqrt{5/32} \\ \sqrt{5/32} \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{5/12} \\ \sqrt{5/12} \\ \sqrt{5/12} \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 3\sqrt{5/68} \\ 2\sqrt{5/68} \\ \sqrt{5/68} \\ 0 \end{bmatrix}$
5	$\mu_1 = \begin{bmatrix} \sqrt{3/5} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{3/8} \\ \sqrt{3/8} \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 2\sqrt{3/17} \\ \sqrt{3/17} \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{1/3} \\ \sqrt{1/3} \\ \sqrt{1/3} \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 3\sqrt{1/6} \\ 2\sqrt{1/6} \\ \sqrt{1/6} \\ 0 \\ 0 \end{bmatrix}$
10	$\mu_1 = \begin{bmatrix} \sqrt{0.55000} \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{0.30556} \\ \sqrt{0.30556} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 2\sqrt{0.11957} \\ \sqrt{0.11957} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{0.22917} \\ \sqrt{0.22917} \\ \sqrt{0.22917} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 3\sqrt{0.04661} \\ 2\sqrt{0.04661} \\ \sqrt{0.04661} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

Table 5: Twenty Conditions Examined when $\delta = 3$

p	1 Variable Shift	2 Variable Shift (Equal)	2 Variable Shift (Unequal)	3 Variable Shift (Equal)	3 Variable Shift (Unequal)
3	$\mu_1 = \begin{bmatrix} \sqrt{2} \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{3/2} \\ \sqrt{3/2} \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 2\sqrt{6/11} \\ \sqrt{6/11} \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \\ \sqrt{2} \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 3\sqrt{3/10} \\ 2\sqrt{3/10} \\ \sqrt{3/10} \end{bmatrix}$
4	$\mu_1 = \begin{bmatrix} \sqrt{15/8} \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{5/4} \\ \sqrt{5/4} \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 2\sqrt{15/32} \\ \sqrt{15/32} \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{5/4} \\ \sqrt{5/4} \\ \sqrt{5/4} \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 3\sqrt{15/68} \\ 2\sqrt{15/68} \\ \sqrt{15/68} \\ 0 \end{bmatrix}$
5	$\mu_1 = \begin{bmatrix} \sqrt{9/5} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{9/8} \\ \sqrt{9/8} \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 2\sqrt{9/17} \\ \sqrt{9/17} \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 3\sqrt{1/2} \\ 2\sqrt{1/2} \\ \sqrt{1/2} \\ 0 \\ 0 \end{bmatrix}$
10	$\mu_1 = \begin{bmatrix} \sqrt{1.65000} \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{0.91667} \\ \sqrt{0.91667} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 2\sqrt{0.35870} \\ \sqrt{0.35870} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} \sqrt{0.68750} \\ \sqrt{0.68750} \\ \sqrt{0.68750} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$	$\mu_1 = \begin{bmatrix} 3\sqrt{0.13983} \\ 2\sqrt{0.13983} \\ \sqrt{0.13983} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

variables that changed. These definitions are similar to the examples shown in Figures 1-3.

MEWMA simulations were conducted using $\delta = 1$ and a one variable shift such that 10,000 out of control signals were obtained. Using $p = 3$, $h = 10.97$, and $\lambda = 0.10$, it was found that the smallest reduced MEWMA statistic correctly identified the variable that changed in 85.24% of the simulations. Using $p = 4$, $h = 12.93$, $\lambda = 0.10$, it was found that the smallest reduced MEWMA statistic correctly identified the variable that changed in 84.73% of the simulations. In addition, using $p = 5$, $h = 14.74$, and $\lambda = 0.10$, showed that the smallest reduced MEWMA statistic correctly identified the variable that changed in 83.84% of the simulations. Using $p = 10$, $h = 22.91$, and $\lambda = 0.10$, it was found that the smallest reduced MEWMA statistic correctly identified the variable that changed in 81.85% of the simulations. Simulations using $\delta = 3$ and a one variable shift produced better results such that when $p = 3, 4, 5$ and 10 , the smallest reduced MEWMA statistic successfully identified the variable that changed in 88.53%, 87.58%, 87.18% and 86.43% of the simulations respectively.

The success rate of one-variable deletion correctly identifying the source of the signal declines as the number of variables shifting increases. For example, using $p = 3$, $h = 10.97$, $\lambda = 0.10$, $\delta = 1$ and an equal-sized two variable shift, the two smallest reduced MEWMA statistics correctly identify the two variables that shifted in 26.89% of the 10,000 simulations. The success rate rapidly declines when three variables shift and the three smallest reduced MEWMA statistics are used to identify the variables that changed. Figures 9 and 10 display the success rates of one-variable deletion when $\delta = 1$ and $\delta = 3$ respectively.

Two-Variable Deletion Analysis

If one variable shifts, the two-variable deletion is considered to be a success if the $(p-1)$ smallest reduced MEWMA statistics correctly identify the variable that changed. If two variables shift, the two-variable deletion is considered to be a success if the smallest reduced MEWMA statistics correctly identify

the two variables that changed. If three variables shift, the two-variable deletion is considered to be a success if the three smallest reduced MEWMA statistics correctly identify the three variables that changed. These definitions are similar to the examples shown in Figures 5-7.

MEWMA simulations were conducted using $\delta = 1$ and a one variable shift such that 10,000 out of control signals were obtained. Using $p = 3$, $h = 10.97$, and $\lambda = 0.10$, it was found that the two smallest reduced MEWMA statistics correctly identified the variable that changed in 87.05% of the simulations. Using $p = 4$, $h = 12.93$, $\lambda = 0.10$, it was found that the three smallest reduced MEWMA statistics correctly identified the variable that changed in 77.58% of the simulations. In addition, using $p = 5$, $h = 14.74$, and $\lambda = 0.10$, showed that the four smallest reduced MEWMA statistics correctly identified the variable that changed in 76.33% of the simulations. Using $p = 10$, $h = 22.91$, and $\lambda = 0.10$, it was found that the nine smallest reduced MEWMA statistics correctly identified the variable that changed in 74.28% of the simulations. Simulations using $\delta = 3$ and a one variable shift produced similar or better results such that when $p = 3, 4, 5$ and 10 , the $(p-1)$ smallest reduced MEWMA statistics successfully identified the variable that changed in 85.61%, 81.77%, 84.61% and 80.38% of the simulations respectively.

The success rate of two-variable deletion correctly identifying the source of the signal decreases when two variables shift. However, the decrease is not as pronounced as the one-variable deletion. For example, using $p = 3$, $h = 10.97$, $\lambda = 0.10$, $\delta = 1$ and an equal-sized two variable shift, the smallest reduced MEWMA statistic correctly identifies the two variables that shifted in 77.02% of the 10,000 simulations. Figures 11 and 12 display the success rates of two variable deletion when $\delta = 1$ and $\delta = 3$ respectively. In addition, there tends to be a slight decrease in the success rate when comparing an unequal shift to an equal shift. The success rate rapidly declines when three variables shift and the three smallest reduced MEWMA statistics are used to identify the variables that changed.

Figure 9: Success Rate of One-Variable Deletion when $\delta = 1$

Figure 10: Success Rate of One-Variable Deletion when $\delta = 3$

GRAPHICAL EXAMINATION OF VARIABLE DELETION WITHIN THE MEWMA

Figure 11: Success Rate of Two-Variable Deletion when $\delta = 1$

Figure 12: Success Rate of Two-Variable Deletion when $\delta = 3$

Conclusion

A general procedure for identifying the variables that contribute to the signal of the MEWMA chart was presented. One-variable deletion correctly identified a one variable shift in 82-89% of the simulations. Two-variable deletion correctly identified a one variable shift in 71-87% of the simulations, an equal-sized two variable shift in 61-81% of the simulations, and an unequal-sized two variable shift in 49-80% of the simulations.

The success rate decreases rapidly when more variables shift than are removed from the MEWMA statistic. However, examining the reduced MEWMA statistics indicated that the criteria employed herein for a successful identification may not immediately identify the variables that contributed to the signal; however, they did lead to a significantly reduced set of variables to search for the cause of the signal.

This study used only one defined covariance matrix such that the correlation between each pair of variables was 0.5. It is suspected that an increase in the success rates would be observed if the correlation between the variables is small. In several of the simulations it was noted that, when a variable would shift, it would drag other variables along with it. This in turn clouded the reduced MEWMA statistics making it more difficult to identify the variable that changed using the previously discussed definitions of a success. Further study is required using different covariance matrices.

In addition, the reported success rates assumed if q-variables shifted, then the corresponding definition of a success was used. Further study is required to examine the success rates using various definitions of a success. One suggestion might be that critical values be established to indicate to the operator that a reduced MEWMA statistic is significantly small. Additional simulations should examine the entire distribution of the reduced MEWMA statistics. Critical values could be obtained by examining the distribution of the reduced MEWMA statistics whose variables had not shifted.

Given the power of today's modern computers, variable deletion could be extended to more than two variables being removed from the calculations. Computers could provide a

sequential method of analysis in which an operator examines one variable deletion results, then two-variable deletion results, three-variable deletion results, etc. Using such a method, it is anticipated that a reasonable success rate for identifying a q-variable shift using a q-variable deletion would be determined. However, this success rate would likely decrease as more variables are added to the process. Additional research is required in this area.

Using variable deletion in conjunction with the MEWMA control chart should enable a user to employ an efficient multivariate control chart with an effective post hoc analysis. In addition, it provides a helpful and easy to understand graphical solution to the problem of identifying which variable(s) contributed to the signal.

References

- Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13, 469-475.
- Alt, F. B. (1984). Multivariate quality control. In *The encyclopedia of statistical sciences*, S. Kotz, N. L. Johnson and C. R. Read, Eds., 110-112. New York, NY: John Wiley.
- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York, NY: Wiley.
- Crossier, R. B. (1988). Multivariate generalization of cumulative sum quality-control schemes. *Technometrics*, 30, 291-303.
- Hotelling, H. (1947). Multivariate quality control illustrated by the air testing of sample bombsights. In *Techniques of statistical analysis*, C. Eisenhart, M. W. Hastay, and W. A. Wallis, Eds., 111-184. New York, NY: McGraw-Hill.
- Jackson, J. E. (1980). Principal components and factor analysis: Part I-principal components. *Journal of Quality Technology*, 12, 201-213.
- Jackson, J. E. (1985). Multivariate quality control. *Communications in Statistics – Theory and Methods*, 14, 2657-2688.
- Jackson, J. E. (1991). *A user's guide to principal components*. New York, NY: John Wiley & Sons, Inc.

GRAPHICAL EXAMINATION OF VARIABLE DELETION WITHIN THE MEWMA

Lowry, C. A., Woodall, W. H., Champ, C. W. & Rigon, S. E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, 34, 46-53.

Lowry, C. A., & Montgomery, D. C. (1995). Review of multivariate control charts. *IIE Transactions*, 27, 800-810.

Mason, R. L., Tracy, N. D., & Young, J. C. (1995). Decomposition of T^2 for multivariate control chart interpretation. *Journal of Quality Technology*, 27, 99-108.

Mason, R. L., Champ, C. W., Tracy, N. D., Wierda, S. J., & Young, J. C. (1997). Assessment of multivariate process control techniques. *Journal of Quality Technology*, 29, 140-143.

Pignatiello, J. J., & Runger, G. C. (1990). Comparison of multivariate CUSUM charts. *Journal of Quality Control*, 22, 173-186.

Wierda, S. J. (1994). Multivariate statistical process control: Recent results and directions for future research. *Statistica Neerlandica*, 48, 147-168.

Woodall, W. H., & Montgomery, D. C. (1999). Research issues and ideas in statistical process control. *Journal of Quality Technology*, 31, 376-386.

Regression Split by Levels of the Dependent Variable

Stan Lipovetsky
GfK Custom Research North America,
Minneapolis, MN

Multiple regression coefficients split by the levels of the dependent variable are examined. The decomposition of the coefficients can be defined by points on the ordinal scale or by levels in the numerical response using the Gifi system of binary variables. This approach permits consideration of specific values of the coefficients at each layer of the response variable. Numerical results illustrate how to identify levels of interpretable regression coefficients.

Key words: Regression model, Gifi system, regression coefficients, levels of response.

Introduction

Interpretation of ordinary least squares (OLS) multiple linear regression with multicollinearity is a well-known problem that has been described in numerous works. The problem is caused by a deteriorating effect that multicollinearity between predictors can produce on OLS coefficients. OLS yields the best aggregate of predictors to fit data – and it is perfect for prediction – but it was not designed to obtain meaningful coefficients for individual predictors in regression (Abraham & Ledolter, 1983; Weisberg, 1985; Andersen & Skovgaard, 2010). Depending on the data, such a model could be useless in analyzing predictor impact on the dependent variable (DV) because multicollinearity yields inflated regression coefficients, pushing them towards large values of both signs. For

example, if multicollinearity yields a negative sign for a presumably useful variable in the model, it is difficult to decide whether it makes sense to increase the value of such a variable to obtain a lift in the output. The techniques for constructing regression models with interpretable coefficients and contributions to the explained variance include ridge regressions (Hoerl & Kennard, 1970, 2000) and various other techniques, particularly: Shapley value regression, logit and multinomial parameterization of coefficients, and models by data gradients (Lipovetsky & Conklin, 2001, 2010c; Lipovetsky, 2009, 2010a, b; Nowakowska, 2010).

The possibility of splitting regression coefficients by the levels of the response variable and studying them separately is considered herein. This will help identify how the obtained coefficients are composed depending on the different values reached by the dependent variable (DV), and how this composition creates the total values of the coefficients. The technique is demonstrated for a DV measured using a numerical and rating scale (such as a Likert-type scale from 1 to 5 or 1 to 10), using the so-called Gifi system of binary multivariables (Gifi, 1990; Michailidis & de Leeuw, 1998; Mair & de Leeuw, 2010), where a variable on a several-point scale can be represented as a set of binary variables – one for each level. For example, a DV on a 5-point scale is presented as the first binary variable with ones in the place of 1s in the original variable and zeroes otherwise, up to the fifth binary variable

Stan Lipovetsky, Ph.D. is Senior Research Director at the GfK Research Center for Excellence, Marketing Sciences. He has numerous publications in multivariate statistics, multiple criteria decision making, econometrics, microeconomics and marketing research. He is a member of the editorial boards of the International Journal of Operations and Quantitative Management, the Journal of Electronic Modeling and the Journal of Model Assisted Statistics and Applications. Email him at: stan.lipovetsky@gfk.com.

where ones represent 5s in the original variable and zeroes otherwise. It is sometimes convenient to consider a fewer number of binary variables, for example, in key dissatisfaction analysis (Conklin, et al., 2004), it is sufficient to use only three binary variables: dissatisfaction (lower levels), neutral (middle), and enhanced values (upper levels). Regressions with interpretable coefficients attained by the split solutions help decision makers and managers understand the results of statistical modeling.

Regression Coefficients by Levels of the DV

A multiple linear regression can be presented as the model

$$y_i = a_0 x_{i0} + a_1 x_{i1} + a_2 x_{i2} + \dots + a_n x_{in} + \varepsilon_i \quad (1)$$

where y_i and x_{ij} are i^{th} observations ($i = 1, \dots, N$) by the DV y and by each j^{th} independent variable x_j ($j = 0, 1, 2, \dots, n$), a_j are coefficients of the regression, including the intercept a_0 related to the identity variable x_0 , and ε_i denotes added random noise. The OLS objective minimizes the squared errors ε_i and yields the solution which, in matrix notation, is:

$$a = (X'X)^{-1} X'y \quad (2)$$

where a denotes the vector of all coefficients of regression (1), X is the design matrix of N by $1+n$ order of all the predictors, prime denotes transposition and vector y is of the N^{th} order.

Formula (2) shows that the regression coefficients are linear combinations of the y values aggregated with the coefficients of the transfer operator $T \equiv (X'X)^{-1} X'$ which depends only on the independent variables. Each j^{th} coefficient a_j is defined as a scalar product of the vector y and the values in the j^{th} row of this matrix T . Therefore, if the vector y is presented as a sum of several sub-vectors then it is possible to obtain the coefficients (2) related to each of these components.

Suppose y is measured in a rating scale of K values, so it can be presented as:

$$y = m_1 d_1 + m_2 d_2 + \dots + m_K d_K, \quad (3)$$

where each d_k ($k = 1, 2, \dots, K$) is a binary vector of the N^{th} order, which has ones in the positions where y_i has the value k , otherwise it consists of zeros. For example, if $y_i = 3$ for $i = 10, 15$ and 18 , then the binary vector d_3 has ones in the same 10^{th} , 15^{th} and 18^{th} places, otherwise zero, and similarly with the other vectors. Such a system of binary variables is called the Gifi system. The constant coefficients m_k in (3) for a Likert scale with ratings from 1 to K coincide with these values, so $m_k = k$. If y is a numerical variable, then it can be divided into several segments by its increasing values, and coefficients m_k represent the mean y values within each segment while the Gifi binary vectors d_k show by 1 and 0 values the particular segment to which each y_i belongs.

Substituting (3) into (2) yields the decomposition of the regression coefficients by the levels of y :

$$\begin{aligned} a &= (X'X)^{-1} X'(m_1 d_1 + m_2 d_2 + \dots + m_K d_K) \\ &= m_1 (X'X)^{-1} X'd_1 + m_2 (X'X)^{-1} X'd_2 \\ &\quad + \dots + m_K (X'X)^{-1} X'd_K \end{aligned} \quad (4)$$

Each matrix product $(X'X)^{-1} X'd_k$ in (4) is the vector of regression coefficients of the binary variable d_k by all the predictors x . It can also be described as the Fisher discriminator of the observations' assignment to each k^{th} segment of the data, thus the total regression coefficients are presented as the linear combination of these discriminators. For each particular level of y -values the coefficients of the Gifi response regressions can be denoted as:

$$b^{(k)} \equiv (X'X)^{-1} X'd_k, \quad (5)$$

and the items in decomposition (4) of the total vector of regression coefficients by the coefficients defined on each y -level are:

$$\begin{aligned} a^{(k)} &= m_k b^{(k)} \\ &= m_k (X'X)^{-1} X'd_k \end{aligned} \quad (6)$$

It is also possible to consider regression results split by the independent variables, but the analysis becomes more complicated (Lipovetsky & Conklin, 2005).

Regression coefficients (2) or (4) can be presented as a sum of the by-level coefficients (6):

$$a = a^{(1)} + a^{(2)} + \dots + a^{(K)} \quad (7)$$

This result is very useful for practical applications of regression modeling because it permits consideration of the increments which can be reached specifically at any level of the dependent variable. For example, in marketing research studies on satisfaction with a product or service, it is useful to consider the regression subsets related to the lower levels of dissatisfaction and upper levels of enhanced satisfaction. The cumulative subtotal of coefficients can also be obtained by adding the needed vectors (7): its application to the multicollinearity problem is discussed next using numerical examples.

Numerical Examples

Consider a marketing research project with 623 respondents evaluating their satisfaction with a bank on a Likert scale from 1 (worst value) to 5 (best value). The variables are: y , overall satisfaction; x_1 , customer service regarding checking account; x_2 , explanations of features; x_3 , kept informed of changes; x_4 , convenient branch locations; x_5 , convenient ATM locations; x_6 , error free checking; x_7 , representative solves problems; and x_8 , clear comprehensive statements. The constructed OLS regression model (1)-(2) is:

$$\begin{aligned} y &= 1.664 + 0.155x_1 + 0.150x_2 \\ &+ 0.123x_3 + 0.005x_4 - 0.012x_5 \\ &+ 0.044x_6 + 0.137x_7 - 0.012x_8 \end{aligned} \quad (8)$$

The main impact on overall satisfaction comes from predictors x_1 , x_2 , x_3 and x_7 . Despite the positive impact that can be assumed for each driver on satisfaction, which is supported by positive pair correlations of each x with y , multicollinearity makes x_4 and x_6 negligibly small coefficients and yields a negative influence on both x_5 and x_8 .

The five Gifi binary regressions for each level of overall satisfaction estimated by (5) are presented in the columns of Table 1. It is evident that the predictors have mixed coefficients for all levels of y , except the top level ($k = 5$), which has all positive coefficients of regression (the negative is the intercept).

Multiplying vectors (5) in the columns of Table 1 by the values $m_k = k$ transforms them into the components (6) of the original regression. These coefficients (6) and their total (7) are shown in Table 2. The coefficients of the last column in Table 2 coincide with the coefficients of the OLS model (8). In contrast to these OLS coefficients with both signs, the $k = 5$ model yields all positive coefficients.

Another useful way to consider the regression coefficients by splitting the cumulative levels of the response is shown in Table 3. It is clear from relation (7) that it is possible to consider subtotals of the split to lower and upper levels. Table 3 presents pairs of the models of the first level versus all other levels (columns denoted as 1 vs. 2:5), two lower and three upper levels (columns 1:2 vs. 3:5), three lower and two upper levels (columns 1:3 vs. 4:5), then four lower versus one upper level (columns 1:4 vs. 5), and finally the total regression by all the levels together (1)-(2). The last row in Table 3 presents the coefficients of multiple determination, R^2 , well-known as a convenient characteristic of quality of the regression model. The sum of the coefficients in each pair of lower and upper models yields the total OLS coefficients of the last column in Table 3. This is not true with the coefficient of multiple determination, R^2 , which is not a linear function of the DV values.

It is observed that a sum of two R^2 values in the paired columns in Table 3 can be higher or lower but not equal to $R^2=0.297$ of the total model. Table 3 also shows that the expected signs of the predictors' relation to the

REGRESSION SPLIT BY LEVELS OF THE DEPENDENT VARIABLE

dependent variable are given only by the upper level of overall satisfaction. Similar results are observed in various data sets.

Conclusion

Decomposition of multiple regression coefficients by the levels of the dependent variable was considered using the Gifi system of binary variables. The coefficients' split by the levels of the response variable can be easily performed with any software for ordinary least squares regression. The results of this decomposition help identify the subsets of the coefficients not distorted by multicollinearity and find an adequate interpretation of the regression coefficients that will be useful for managerial decisions.

References

- Abraham, B., & Ledolter, J. (1983). *Statistical methods for forecasting*. New York, NY: Wiley.
- Andersen, P. K., & Skovgaard, L. T. (2010). *Regression with linear predictors*. New York, NY: Springer.
- Conklin, M., Powaga, K., & Lipovetsky, S. (2004). Customer satisfaction analysis: Identification of key drivers. *European Journal of Operational Research*, 154, 819-827.
- Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, England: Wiley.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12, 55-67.

Table 1: Coefficients of Gifi Response Regressions (5)

Coefficients	$b^{(1)}$	$b^{(2)}$	$b^{(3)}$	$b^{(4)}$	$b^{(5)}$
$b_0^{(k)}$	0.156	0.175	0.729	0.727	-0.787
$b_1^{(k)}$	-0.008	-0.030	-0.017	0.003	0.052
$b_2^{(k)}$	-0.009	-0.027	0.012	-0.054	0.079
$b_3^{(k)}$	-0.012	0.000	-0.046	0.018	0.040
$b_4^{(k)}$	0.010	0.004	-0.020	-0.015	0.021
$b_5^{(k)}$	0.006	0.014	-0.012	-0.029	0.021
$b_6^{(k)}$	-0.006	0.007	-0.036	0.031	0.004
$b_7^{(k)}$	-0.017	-0.014	-0.027	0.024	0.033
$b_8^{(k)}$	0.005	0.010	0.005	-0.051	0.030

Table 2: Coefficients of Regression Split by Levels of Response (6)

Coefficients	$a^{(1)}$	$a^{(2)}$	$a^{(3)}$	$a^{(4)}$	$a^{(5)}$	Total a
$a_0^{(k)}$	0.156	0.350	2.188	2.907	-3.937	1.664
$a_1^{(k)}$	-0.008	-0.061	-0.051	0.013	0.261	0.155
$a_2^{(k)}$	-0.009	-0.054	0.035	-0.216	0.396	0.150
$a_3^{(k)}$	-0.012	0.000	-0.139	0.074	0.200	0.123
$a_4^{(k)}$	0.010	0.008	-0.061	-0.059	0.107	0.005
$a_5^{(k)}$	0.006	0.028	-0.037	-0.115	0.106	-0.012
$a_6^{(k)}$	-0.006	0.015	-0.108	0.125	0.018	0.044
$a_7^{(k)}$	-0.017	-0.027	-0.082	0.096	0.167	0.137
$a_8^{(k)}$	0.005	0.021	0.016	-0.205	0.151	-0.012

Table 3: Coefficients of Regression with the DV Split into Two Levels

	1 vs.2:5		1:2 vs. 3:5		1:3 vs. 4:5		1:4 vs. 5		1:5
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Total
a_0	0.156	1.508	0.506	1.158	2.694	-1.030	5.602	-3.937	1.664
a_1	-0.008	0.163	-0.069	0.224	-0.120	0.275	-0.106	0.261	0.155
a_2	-0.009	0.160	-0.064	0.214	-0.029	0.180	-0.245	0.396	0.150
a_3	-0.012	0.136	-0.012	0.135	-0.151	0.274	-0.077	0.200	0.123
a_4	0.010	-0.005	0.017	-0.013	-0.043	0.048	-0.102	0.107	0.005
a_5	0.006	-0.018	0.034	-0.046	-0.003	-0.009	-0.118	0.106	-0.012
a_6	-0.006	0.050	0.008	0.035	-0.100	0.143	0.026	0.018	0.044
a_7	-0.017	0.154	-0.044	0.181	-0.126	0.263	-0.030	0.167	0.137
a_8	0.005	-0.017	0.026	-0.037	0.042	-0.053	-0.163	0.151	-0.012
R^2	0.070	0.289	0.139	0.288	0.184	0.280	0.090	0.166	0.297

Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42, 80-86.

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17, 319-330.

Lipovetsky, S., & Conklin, M. (2005). Regression by data segments via discriminant analysis. *Journal of Modern Applied Statistical Methods*, 4, 63-74.

Lipovetsky, S. (2009). Linear regression with special coefficient features attained via parameterization in exponential, logistic, and multinomial-logit forms. *Mathematical and Computer Modelling*, 49, 1427-1435.

Lipovetsky, S. (2010a). Enhanced ridge regressions. *Mathematical and Computer Modelling*, 51, 338-348.

Lipovetsky, S. (2010b). Meaningful regression coefficients built by data gradients. *Advances in Adaptive Data Analysis*, 2, 451-462.

Lipovetsky, S., & Conklin, M. (2010c). Meaningful regression analysis in adjusted coefficients Shapley value model. *Model Assisted Statistics and Applications*, 5, 251-264.

Mair, P., & de Leeuw, J. (2010). A general framework for multivariate analysis with optimal scaling: The R package aspect. *Journal of Statistical Software*, 32, 1-23.

Michailidis, G., & de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. *Statistical Science* 13: 307-336.

Nowakowska, E. (2010). Modeling in a multicollinear setup: Determinants of SVR advantage. *Model Assisted Statistics and Applications*, 5, 219-233.

Weisberg, S. (1985). *Applied linear regression*. New York, NY: Wiley.

Testing the Population Coefficient of Variation

Shipra Banik
Independent University,
Dhaka, Bangladesh

B. M. Golam Kibria
Florida International University
University Park, FL

Dinesh Sharma
James Madison University
Harrisonburg, VA

The coefficient of variation (CV), which is used in many scientific areas, measures the variability of a population relative to its mean and standard deviation. Several methods exist for testing the population CV. This article compares a proposed bootstrap method to existing methods. A simulation study was conducted under both symmetric and skewed distributions to compare the performance of test statistics with respect to empirical size and power. Results indicate that some of the proposed methods are useful and can be recommended to practitioners.

Key words: Coefficient of variation, simulation, size, power of a test, symmetric distribution, skewed distribution.

Introduction

The coefficient of variation (CV), which is the ratio of the standard deviation to the mean, was first introduced by Karl Pearson in 1896. This dimensionless relative measure of dispersion has widespread applications in many disciplines. Researchers have used CV to: measure the risk of a stock (Miller & Karson, 1977), to assess the strength of ceramics (Gong & Li, 1999), to assess homogeneity of bone test samples produced from a particular method (Hamer, et al., 1995), in wildlife studies (Dodd & Murphy, 1995), in dose-response studies (Creticos, et al., 2002) and in uncertainty analyses of fault tree analysis (Ahn, 1995). Nair and Rao (2003)

provided a brief survey of recent applications of CV in business, climatology, engineering and other fields.

The coefficient of variation is presented in virtually all introductory statistics texts, primarily as a descriptive measure; inferential methods regarding population CVs are typically missing in these textbooks. To make an inference regarding a population CV, assumptions regarding the population distribution and knowledge of the distributional properties of the sample CV are needed. Hendricks and Robey (1936) studied the distribution of the sample CV and showed that it can be approximated by a function defined on a positive real line, which depends on the standard normal moment of order $n - 1$ about some well-defined point, where n is the sample size. Iglewicz (1967) derived the exact distribution for a sample CV, when the sample is drawn from a normal population. This exact distribution assumed that the chance of obtaining a non-positive sample mean is negligible and, hence, is not useful for inferential purposes.

McKay (1932) gave an approximation of the distribution of a statistic derived from a sample CV based on the Chi-squared distribution. This approximation was determined to be very accurate if $CV \leq 0.33$ (Pearson, 1932; Iglewicz, 1967) and reasonably accurate when

Shipra Banik is an Associate Professor in the School of Engineering and Computer Science. Email: banik@secs.iub.edu.bd. B. M. Golam Kibria is an Associate Professor in the Department of Mathematics and Statistics. He is the coordinating editor for the *Journal of Probability and Statistical Science*. He is an elected fellow of the *Royal Statistical Society* and the *International Statistical Institute*. Email: kibriag@fiu.edu. Dinesh Sharma is an Assistant Professor in the Department of Mathematics and Statistics. Email: sharmadr@jmu.edu.

TESTING THE POPULATION COEFFICIENT OF VARIATION

$0.33 \leq CV \leq 0.67$ (Miller, 1991). The exact distribution of the sample CV is difficult to obtain when the population distribution is not normal. Due to the limited development related to the exact distribution of sample CV for non-normal populations, inferences regarding population CVs did not receive much attention until Sharma and Krishna (1994) developed the asymptotic distribution of the sample inverse coefficient of variation (ICV) without making an assumption about the population distribution; they obtained a confidence interval for the CV by inverting the proposed confidence interval. Curto and Pinto (2009) developed an asymptotic distribution of a sample CV in the case of non-iid (independent and identically distributed) random variables.

Various methods for constructing confidence intervals on CV have recently appeared in the literature (Amiri & Zwanzig, 2010; Carto & Pinto, 2009; Banik & Kibria, 2011). Banik and Kibria (2011) conducted a simulation study to compare the performance of various confidence intervals suggested in the literature. However, despite its widespread use in a wide range of disciplines, tests of hypotheses on CVs do not appear to be of interest to statisticians in general. Although some test statistics have been suggested, there is limited information available regarding the performance of these tests. Moreover, many of these tests are based on normal theory; however, real life data frequently follow right-skewed distributions, particularly when sample sizes are small (Baklizi & Kibria, 2008; Shi & Kibria, 2007; Banik & Kibria, 2009; Almonte & Kibria, 2009).

This article compares the size and the power of some existing tests and their bootstrap versions when data are from both normal and positively skewed distributions. The tests compared were developed based on the sampling distribution of a sample CV due to McKay (1932), Hendricks and Robey (1936), Miller (1991), Sharma and Krishna (1994) and Curto and Pinto (2009).

Test Statistics for Testing Population CV

Let $X_1, X_2, \dots, X_n$ represent a random sample of size n from a normal population with

mean μ and SD σ so that $\gamma = \sigma / \mu$ is the population CV. When the distribution is unknown, the parameters μ and σ are estimated from the observed data. The estimated CV is then defined as $\hat{\gamma} = s / \bar{X}$ where $\bar{X}$ and s are the sample mean and sample standard deviation respectively. The test hypotheses are:

$$\begin{aligned} H_0 : \gamma &= \gamma_0 \\ \text{vs.} \\ H_a : \gamma &= \gamma_1 \end{aligned} \quad (1)$$

where $\gamma_1 = \gamma_0 + c$, c is a positive constant and the difference between γ_0 and the true value of the population CV. Because the right skewed distribution is of interest, the upper tailed test was selected. However, the lower tail test may be used by setting $c < 0$. The size of a test can be estimated by setting $c = 0$. Several test statistics have been suggested for testing the hypotheses in (1).

The t -Statistic

Hendricks and Robey (1936) studied the distribution of a sample CV when the sample is drawn from a normal distribution. Koopmans, et al (1964) and Igelewicz (1967) reviewed the relevant literature and proposed the following t -statistic for testing γ for a normal distribution:

$$t_0 = \frac{(\hat{\gamma} - \gamma_0)}{S_{\hat{\gamma}} \sim t_{(n-1)}} \quad (2)$$

where $\hat{\gamma}$ is the sample CV, $S_{\hat{\gamma}} = \hat{\gamma} / \sqrt{2n}$. The null hypothesis is rejected at the α level of significance if $t_0 > t_{(n-1), \alpha}$ where $t_{(n-1), \alpha}$ is the upper $(\alpha)^{\text{th}}$ percentile from a t -distribution with $(n - 1)$ degrees of freedom.

McKay's Statistic

McKay (1932) proposed the following test statistic for testing γ .

$$Mc = \left(1 + \gamma_0^{-1}\right) \left(\frac{n\hat{\gamma}^2}{1 + \hat{\gamma}^2}\right) \sim \chi_{(n-1)}^2 \quad (3)$$

Where $\hat{\gamma}$ is the sample CV. The hull hypothesis is rejected at the α level of significance if $Mc > \chi^2_{(n-1), \alpha}$ where $\chi^2_{(n-1), \alpha}$ is the upper $(\alpha)^{th}$ percentile from a Chi-square distribution with $(n - 1)$ degrees of freedom.

Miller's Statistic

Miller (1991) provided an asymptotic distribution of the sample CV which can reasonably be assumed to be normal if the parent population is normal. They proposed the following test statistic:

$$MiL = (\hat{\gamma} - \gamma_0) / S_{\hat{\gamma}_M} \sim Z(0,1), \quad S_{\hat{\gamma}_M} = \sqrt{(\hat{\gamma}^4 + 0.5\hat{\gamma}^2) / n} \quad (3)$$

Sharma and Krishna's Statistic

Sharma and Krishna's (1994) statistic, which is based on the sampling distribution of ICV, is given by

$$SK = n(\hat{\gamma} - \gamma_0) \sim (1 / Z(0,1)) \quad (4)$$

As noted, this has the advantage of relieving the normality assumption.

Curto and Pinto's Statistic

Curto and Pinto (2009b) proposed a test statistics for non-iid random variables, that is, autocorrelated and heteroskedastic random variables. Their test statistic is given by:

$$CP = (\hat{\gamma} - \gamma_0) / SE(\hat{\gamma}) \quad (5)$$

where

$$SE(\hat{\gamma}) \equiv \sqrt{V_{GMM} / n},$$

$$V_{GMM} = \frac{\partial f(\theta)}{\partial \theta} \Sigma \frac{\partial f(\theta)}{\partial \theta'},$$

and

$$\frac{\partial f(\theta)}{\partial \theta'} = \begin{bmatrix} -\sigma / \mu^2 \\ \frac{1}{2\sigma\mu} \end{bmatrix}$$

To estimate the asymptotic variance, an estimator for $\partial f(\theta) / \partial \theta'$ may be obtained by substituting into $\hat{\theta}$ and a heteroscedasticity and autocorrelation consistent (HAC) estimator $\hat{\Sigma}$ may be obtained by using Newey and West's (1987) procedure:

$$\hat{\Sigma} = \hat{\Omega}_0 + \sum_{j=1}^m \omega(j, m) (\hat{\Omega}_j + \hat{\Omega}_j'),$$

$$\hat{\Omega}_j = \frac{1}{T} \sum_{t=j+1}^T \phi(X_t, \hat{\theta}) \phi(X_t, \hat{\theta})',$$

$$\phi(X_t, \hat{\theta}) = \begin{bmatrix} X_t - \hat{\mu} \\ (X_t - \hat{\mu})^2 - \hat{\sigma}^2 \end{bmatrix},$$

and

$$\omega(j, m) = 1 - \frac{j}{m+1},$$

where m is the truncated lag that must satisfy the condition m/T .

Proposed Bootstrap Test Statistics for Testing Population CV

Bootstrap, introduced by Efron (1979), is a commonly used computer-based non-parametric tool that does not require assumptions regarding an underlying population and can be applied in a variety of situations. The accuracy of the bootstrap depends on the number of bootstrap samples. If the number of bootstrap samples is large enough, statistics may be more accurate. The number of bootstrap samples is typically between 1,000 and 2,000 because accuracy depends on the size of the samples (Efron & Tibshirani, 1993). This article proposes bootstrap test statistics for testing a population CV. An extensive array of different bootstrap methods are summarized as: Let $X^{(*)} = X_1^{(*)}, X_2^{(*)}, \dots, X_n^{(*)}$, where the i^{th} sample is denoted $X^{(i)}$ for $i = 1, 2, \dots, B$, and B is the number of bootstrap samples. The bootstrap estimate of CV for the B samples is $CV_{(i)}^*$.

TESTING THE POPULATION COEFFICIENT OF VARIATION

Non-Parametric Bootstrap Statistic

First, compute the CV for all bootstrap samples, then order the sample CVs of each bootstrap sample as:

$$CV_{(1)}^* \leq CV_{(2)}^* \dots \leq CV_{(B)}^*$$

The test statistic for testing hypotheses (1) is the t-statistic defined in (2) but the $(1-\alpha)$ sample quantile of the bootstrap samples, $CV_{[(1-\alpha)B]}^*$, is used as the upper critical value for the test.

Parametric Bootstrap t-Statistic

The bootstrap version of the t -statistic defined in (2) is given by

$$BT_i^* = \frac{(CV_i^* - \overline{CV})}{\hat{\sigma}_{CV}}, \quad (6)$$

$i = 1, 2, \dots, B,$

where $\hat{\sigma}_{CV} = \frac{1}{B} \sum_{i=1}^B (CV_i^* - \overline{CV})^2$, and

$\overline{CV} = \frac{1}{B} \sum_{i=1}^B CV_i^*$ is the mean of the bootstrap sample CVs. The $(1-\alpha)^{th}$ quantile of the bootstrap t -statistic in (6) is used as the upper critical value for an α level test.

Miller Bootstrap Statistics: Approach 1

This approach suggests replacing $\hat{\gamma}$ in (3) by $\hat{\gamma}^*$, the sample CV of the bootstrap sample, thus, the following test statistic is proposed:

$$BMiL1 = \frac{\hat{\gamma}^* - \gamma_0}{S_{\hat{\lambda}_M}} \sim Z(0,1), \quad (7)$$

$$S_{\hat{\lambda}_M} = \sqrt{\frac{(\hat{\gamma}^{*4} + 0.5\hat{\gamma}^{*2})}{n}}.$$

Miller Bootstrap Statistics: Approach 2

As noted, the approximate asymptotic normality of the sampling distribution of $\hat{\gamma}$ is based on the assumption that the parent population is normal (Miller, 1991); violation of

the normality assumption may lead to undesirable results. The following bootstrap test is thus proposed:

$$BMiL2 = \frac{\hat{\gamma} - \gamma_0}{S_{\hat{\lambda}_M}} \quad (8)$$

where $S_{\hat{\lambda}_M}$ is as defined in (3). The null hypothesis in (1) is rejected if $BMiL2 > Z_{\alpha/2}^*$, where $Z_{\alpha/2}^*$ is the $(1-\alpha)^{th}$ quantile of

$$Z_i^* = (CV_i^* - \overline{CV}) / \hat{\sigma}_{CV}$$

with

$$\hat{\sigma}_{CV} = \frac{1}{B} \sum_{i=1}^B (CV_i^* - \overline{CV})^2$$

and

$$\overline{CV} = \frac{1}{B} \sum_{i=1}^B CV_i^*.$$

Curto and Pinto Bootstrap Statistic

The following test statistic for bootstrap version of CP is proposed:

$$BCP = (\hat{\gamma}^* - \gamma_0) / SE(\hat{\gamma}^*) \quad (9)$$

where $\hat{\gamma}^*$ is the sample CV of the bootstrap samples and

$$SE(\hat{\gamma}^*) \cong \sqrt{BV_{GMM} / B},$$

$$BV_{GMM} = \frac{\partial f(\theta)}{\partial \theta} \Sigma^* \frac{\partial f(\theta)}{\partial \theta'},$$

$$\hat{\Sigma}^* = \hat{\Omega}_0 + \sum_{j=1}^m \omega(j, m) (\hat{\Omega}_j^* + \hat{\Omega}_j^{*'}),$$

$$\hat{\Omega}_j^* = \frac{1}{B} \sum_{t=j+1}^T \phi(X_t^*, \hat{\theta}) \phi(X_t^*, \hat{\theta})',$$

and

$$\varphi(X_t^*, \hat{\theta}) = \left[\frac{X_t^* - \hat{\mu}}{(X_t^* - \hat{\mu})^2 - \hat{\sigma}^2} \right],$$

where m is the truncated lag that must satisfy the condition m/T . The $(1 - \alpha)^{\text{th}}$ quantile of the bootstrap statistic in (9) is used as the upper critical value for an α level test.

Methodology

Monte Carlo simulation experiments were performed to evaluate the performance of the proposed test statistics in terms of size and power. The main objective is to recommend good test statistics for a population CV based on simulation results. Because a theoretical comparison was not possible, a simulation study was used to compare the size and power performances of the test statistics.

Six different configurations of sample sizes: $n = 10, 20, 30, 50, 100, 200$ were used. Random samples were generated from the $N(2, 1)$ and two skewed distributions namely, Gamma (4, 2) and Log-Normal (2, 0.472). This parameter choice resulted in population CVs close to 0.5 for all selected distributions with varying degree of skewness. Note that non-iid data was not used in the simulation. Although the Curto and Pinto statistic was proposed for non-iid, that is, autocorrelated and heteroscedastic random variables, the focus was to compare the size and power of existing test statistics with the proposed bootstrap statistics for testing population CV when data are generated from symmetric and skewed distributions.

For each combination of sample size and population distribution, 10,000 random samples and 1,500 bootstrap replications were generated. The most common 5% level (α) of significance was used. Empirical sizes and powers for each test were calculated as the fraction of the rejections of the null hypothesis out of 10,000 simulation replications by setting $c = 0.0, 0.04, 0.06, 0.08, 0.10$ and 0.12 in $\gamma_1 = \gamma_0 + c$. The size of the test was obtained by setting $c = 0$. Simulation results are presented in Tables 3.1-3.4.

Results

Table 3.1 shows the estimated sizes of the selected test statistics for all three distributions. The row entries represent the proportion of times H_0 was rejected at $\alpha = 0.05$ under H_0 . If a procedure is significantly above the nominal level or significantly above the level of some other procedure, there may be a question about the seriousness of the degree of non-robustness. Rejection rates significantly below the nominal level are not of interest. Such deviations are not problematic for Type I errors and power can be evaluated separately. To gauge the adequacy of robustness in controlling Type I errors, several standards have been used in the past. Cochran (1954) suggested the general guideline of an upper limit of 0.06 for tests run at the 0.05 level. Bradley (1978) considered a liberal criterion of robustness in which he argued that no test should be considered robust if the true Type I error rate exceeds 1.5α ; meaning, that an $\alpha = 0.05$ would require an actual limit of 0.075. Finally Conover, et al. (1981) used a more liberal approach and suggested that a test is non-robust if the Type I error rate exceeds 2α .

From the data shown in Table 3.1 it is clear that none but the *CP* procedure is most conservative. Its size is smaller than the nominal size of 0.05 for all sample size for the normal and Gamma distributions, and for all $n > 10$ for the log-normal distribution. For the Gamma distribution, however, all of the tests suffer from size distortion when $n < 200$.

When the underlying distribution is normal all of the procedures, except the *SK* test satisfy Cochran's 0.06 limit (and hence Bradley's 0.075 and Conover, et al.'s 0.1 limit), particularly when $n > 20$. It is noteworthy that the *SK* procedure does not satisfy Cochran's limit for the normal distribution and no clear superiority of one test is apparent for the log-normal distribution. However, the *t*-test, *SK* and *NB* procedures appear to be clearly non-inferior in controlling Type I errors, especially when $n \geq 100$, as their estimated type I error rate is either very close to or exceeds Conover, et al.'s 0.1 limit. In general, the bootstrap versions of the *Mil* and *CP* tests are slightly more conservative than their respective non-bootstrap counterparts for all three distributions and all tests have reasonable size properties when data are

generated from a normal distribution as opposed to the two skewed distributions.

The estimated powers of the test statistics for the normal, Gamma and log-normal distributions are presented in Tables 3.2, 3.3 and 3.4 respectively. The first column provides values of c , which is the difference between γ_0 and the true value of population CV. The entries in the columns 3-12 represent the proportion of times H_0 was rejected at $\alpha = 0.5$ under H_I . With few exceptions, Miller's procedure appears to be the most powerful under a normal distribution, while the BMil2 shows some advantages over other tests under a Gamma distribution: no clear pattern of dominance of one test is visible for the log-normal distribution. When $c = 0.04$ all tests considered have very low power (maximum power = 0.48). As expected, power increases with c . Most of the tests have reasonable power when $c = 0.12$ and sample size is low, such as $n = 30$ for the normal and Gamma distributions. A comparison of results presented in Table 3.2 reveals that the Curto and Pinto's test and its bootstrap version, BCP, are both relatively more powerful under a Gamma distribution as well as a log-normal distribution compared to the normal case. In addition, the parametric bootstrap test also shows a clear pattern of improvement in power over the normal case.

Conclusion

This article considered five existing test statistics and five bootstrap versions of three of the tests for testing a population CV under various experimental conditions. Because a theoretical comparison is not possible, a simulation study was conducted to compare the performance of the test statistics. Results indicate that all of the test statistics suffer from size distortion, particularly when data is from either a Gamma or a log-normal distribution and $n \leq 50$. None of the tests is recommended if c , the difference between the hypothesized and true value of the population CV is too small, that is, if $c < 0.06$. Although Miller's test appears to be the most powerful under a normal distribution, its bootstrap version, BMil2, shows some advantages over the other tests for Gamma

distributions. Sharma and Krishna's test is most powerful for the normal distribution. For a definite statement regarding the performance of the test statistics, additional simulations under variety of experimental conditions are required. It is hoped that the results from this study will be useful to different applied researchers and practitioners who are interested to test a population CV for symmetric and skewed populations.

References

- Ahn, Kwang-II. (1995). Use of the coefficient of variation for uncertainty analysis in fault tree analysis. *Reliability Engineering and System Safety*, 47(3), 229-230.
- Almonte, C., & Kibria, B. M. G. (2009). On some classical, bootstrap and transformation confidence intervals for estimating the mean of an asymmetrical population. *Model Assisted Statistics and Applications*, 4(2), 91-104.
- Amiri, S., & Zwanzig, S. (2010). An improvement of the nonparametric bootstrap test for the comparison of the coefficient of variations. *Communication in Statistics-Simulation and Computation*, 39(9), 1726-1734.
- Banik, S., & Kibria, B. M. G. (2009). On some discrete distributions and their applications with real life data. *Journal of Modern Applied Statistical Methods*, 8(3), 423-447.
- Banik, S., & Kibria, B. M. G. (2011). Estimating the population coefficient of variation by confidence intervals. *Communications in Statistics-Simulation and Computation*, 40(8), 1236-1261.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10, 417-451.
- Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23, 351-361.

Table 3.1: Estimated Type I Error Rates for Various Statistical Tests

Distribution	Tests for CV*	n					
		10	20	30	50	100	200
Normal (2, 1)	t-test	0.0300	0.0453	0.0473	0.0580	0.0580	0.0593
	McKay	0.0613	0.0607	0.0480	0.0500	0.0380	0.0480
	MiL	0.0793	0.0747	0.0553	0.0580	0.0427	0.0547
	SK	0.0980	0.0867	0.0640	0.0813	0.0753	0.0700
	CP	0.0107	0.0180	0.0160	0.0200	0.0233	0.0287
	NB	0.0333	0.0378	0.0407	0.0520	0.0527	0.0593
	PB	0.0393	0.0480	0.0473	0.0580	0.0500	0.0427
	BMiL1	0.0667	0.0653	0.0507	0.0520	0.0407	0.0493
	BMiL2	0.0740	0.0667	0.0520	0.0547	0.0373	0.0520
	BCP	0.0080	0.0120	0.0127	0.0200	0.0213	0.0253
Gamma (4, 2)	t-test	0.0060	0.0140	0.0240	0.0427	0.0420	0.0520
	McKay	0.0173	0.0207	0.0253	0.0333	0.0287	0.0320
	MiL	0.0247	0.0280	0.0333	0.0427	0.0327	0.0387
	SK	0.0367	0.0307	0.0420	0.0633	0.0507	0.0580
	CP	0.0260	0.0240	0.0360	0.0367	0.0313	0.0293
	NB	0.0187	0.0247	0.0333	0.0427	0.0420	0.0547
	PB	0.0173	0.0207	0.0347	0.0387	0.0420	0.0520
	BMiL1	0.0080	0.0080	0.0167	0.0280	0.0247	0.0280
	BMiL2	0.0367	0.0280	0.0380	0.0367	0.0327	0.0387
	BCP	0.0400	0.0253	0.0427	0.0307	0.0307	0.0293
Log-Normal (2, 0.4720)	t-test	0.0160	0.0287	0.0480	0.0607	0.0953	0.1033
	McKay	0.0313	0.0367	0.0480	0.0487	0.0720	0.0707
	MiL	0.0413	0.0433	0.0560	0.0607	0.0793	0.0767
	SK	0.0487	0.0487	0.0613	0.0827	0.1120	0.1127
	CP	0.0527	0.0440	0.0380	0.0280	0.0293	0.0213
	NB	0.0200	0.0367	0.0593	0.0707	0.1007	0.1240
	PB	0.0467	0.0407	0.0733	0.0753	0.0973	0.0973
	BMiL1	0.0160	0.0193	0.0267	0.0347	0.0573	0.0593
	BMiL2	0.0560	0.0487	0.0733	0.0693	0.0813	0.0720
	BCP	0.0793	0.0533	0.0533	0.0347	0.0313	0.0173

*Notes: t, t statistic; McKay, McKay; MiL, Miller; SK, Sharma and Krishna; CP, Curto and Pinto; NB, Non-parametric bootstrap; PB, Parametric bootstrap; BMiL1, Miller bootstrap 1; BMiL2, Miller bootstrap 2; BCP, Bootstrap Curto and Pinto

TESTING THE POPULATION COEFFICIENT OF VARIATION

Table 3.2: Estimated Power of Various Tests for the Normal (2, 1) Distribution

c	n	Tests for CV*									
		t	McKay	MiL	SK	CP	NB	PB	BMiL1	BMiL2	BCP
0.04	10	0.0427	0.1093	0.1340	0.1307	0.0233	0.0693	0.0413	0.1087	0.1067	0.0100
	20	0.1047	0.1467	0.1647	0.1487	0.0513	0.0713	0.1187	0.1433	0.1587	0.0407
	30	0.1327	0.1447	0.1593	0.1487	0.0680	0.1420	0.1367	0.1427	0.1540	0.0620
	50	0.2087	0.2020	0.2193	0.2313	0.1187	0.2313	0.2267	0.2007	0.2267	0.1307
	100	0.2807	0.2400	0.2987	0.2833	0.1727	0.3053	0.2567	0.2447	0.2373	0.1500
	200	0.4773	0.4187	0.4840	0.4587	0.3380	0.4453	0.4327	0.4180	0.3787	0.2813
0.06	10	0.0620	0.1413	0.1653	0.1467	0.0273	0.1113	0.0493	0.1393	0.1333	0.0107
	20	0.1527	0.2020	0.2273	0.1827	0.0773	0.1473	0.1773	0.1953	0.2313	0.0787
	30	0.2013	0.2327	0.2553	0.2107	0.1107	0.1913	0.2107	0.2240	0.2453	0.1053
	50	0.2793	0.2787	0.2973	0.2913	0.1827	0.2393	0.2693	0.2713	0.2820	0.1607
	100	0.5047	0.4720	0.5187	0.4833	0.3787	0.4733	0.4733	0.4660	0.4533	0.3427
	200	0.7367	0.6813	0.7400	0.6873	0.6193	0.7133	0.7647	0.6800	0.7280	0.6587
0.08	10	0.0847	0.2007	0.2287	0.1920	0.0480	0.1793	0.1173	0.1847	0.2187	0.0393
	20	0.1947	0.2547	0.2800	0.2067	0.1220	0.2247	0.2240	0.2373	0.2820	0.1287
	30	0.2827	0.3280	0.3493	0.2667	0.1760	0.2827	0.2720	0.3147	0.3300	0.1500
	50	0.4373	0.4440	0.4667	0.4240	0.3107	0.4020	0.4040	0.4273	0.4320	0.2627
	100	0.7000	0.6753	0.7207	0.6527	0.5727	0.7067	0.6900	0.6680	0.6760	0.5513
	200	0.9240	0.9053	0.9327	0.8913	0.8767	0.9140	0.9240	0.9040	0.9100	0.8727
0.10	10	0.1207	0.2647	0.2880	0.2313	0.0720	0.1593	0.2193	0.2487	0.3100	0.0960
	20	0.3040	0.3807	0.4033	0.2980	0.1753	0.3173	0.3140	0.3607	0.3900	0.1613
	30	0.4293	0.4827	0.5013	0.3900	0.3033	0.4367	0.4547	0.4627	0.5027	0.3053
	50	0.6087	0.6293	0.6440	0.5587	0.4607	0.6167	0.5553	0.6100	0.6033	0.3940
	100	0.8587	0.8507	0.8560	0.8047	0.7720	0.8540	0.8633	0.8433	0.8593	0.7793
	200	0.9880	0.9867	0.9873	0.9787	0.9773	0.9900	0.9907	0.9847	0.9873	0.9793
0.12	10	0.1580	0.3253	0.3500	0.2507	0.0973	0.1887	0.1700	0.2900	0.3180	0.0673
	20	0.4000	0.4887	0.5007	0.3680	0.2707	0.4333	0.4227	0.4607	0.4960	0.2580
	30	0.5467	0.5973	0.6173	0.4640	0.4133	0.5140	0.5213	0.5793	0.5880	0.3533
	50	0.7453	0.7627	0.7733	0.6853	0.6260	0.7240	0.7233	0.7493	0.7520	0.5740
	100	0.9507	0.9480	0.9513	0.9180	0.9173	0.9327	0.9420	0.9420	0.9420	0.9040
	200	1.0000	1.0000	1.0000	0.9987	0.9993	1.0000	1.0000	1.0000	1.0000	0.9993

*Notes: t, t statistic; McKay, McKay; MiL, Miller; SK, Sharma and Krishna; CP, Curto and Pinto; NB, Non-parametric bootstrap; PB, Parametric bootstrap; BMiL1, Miller bootstrap 1; BMiL2, Miller bootstrap 2; BCP, Bootstrap Curto and Pinto

Table 3.3: Estimated Power of Various Tests for the Gamma (4, 2) Distribution

c	n	Tests for CV*									
		t	McKay	MiL	SK	CP	NB	PB	BMiL1	BMiL2	BCP
0.04	10	0.0107	0.0493	0.0673	0.0660	0.0720	0.0280	0.0247	0.0200	0.0653	0.0680
	20	0.0560	0.0787	0.0993	0.0840	0.0873	0.0787	0.0780	0.0553	0.1093	0.0987
	30	0.0920	0.0993	0.1220	0.1033	0.1080	0.1160	0.1167	0.0787	0.1367	0.1213
	50	0.1353	0.1313	0.1447	0.1553	0.1413	0.1573	0.1327	0.1093	0.1340	0.1267
	100	0.2553	0.2127	0.2333	0.2553	0.2207	0.2493	0.2627	0.1987	0.2420	0.2253
	200	0.4427	0.3760	0.4000	0.4200	0.3933	0.4407	0.4187	0.3700	0.3667	0.3673
0.06	10	0.0240	0.0907	0.1073	0.0933	0.0927	0.0393	0.1253	0.0467	0.1767	0.1607
	20	0.0740	0.1047	0.1280	0.0893	0.1227	0.0853	0.1900	0.0787	0.1320	0.1287
	30	0.1640	0.1887	0.2033	0.1713	0.1893	0.1600	0.2080	0.1540	0.2287	0.2193
	50	0.2807	0.2780	0.2993	0.2907	0.2673	0.2667	0.3040	0.2493	0.3113	0.2840
	100	0.4560	0.4167	0.4400	0.4360	0.4253	0.4753	0.4473	0.3933	0.4220	0.4060
	200	0.7473	0.6920	0.7067	0.6967	0.7160	0.7387	0.7820	0.6773	0.7460	0.7573
0.08	10	0.0300	0.1273	0.1540	0.1167	0.1267	0.0740	0.0740	0.0713	0.1673	0.1413
	20	0.1513	0.2087	0.2333	0.1673	0.2053	0.2300	0.1513	0.1720	0.2093	0.1793
	30	0.2347	0.2633	0.2887	0.2173	0.2693	0.2227	0.2760	0.2293	0.3240	0.3000
	50	0.4213	0.4293	0.4460	0.4000	0.4107	0.4240	0.4507	0.3840	0.4620	0.4253
	100	0.7180	0.7033	0.7120	0.6860	0.7053	0.7340	0.7047	0.6880	0.6987	0.6720
	200	0.9327	0.9140	0.9167	0.8987	0.9193	0.9373	0.9247	0.9053	0.9073	0.9073
0.10	10	0.0560	0.1780	0.2013	0.1433	0.1607	0.0900	0.1093	0.1187	0.2147	0.1720
	20	0.2320	0.3227	0.3473	0.2267	0.3000	0.3160	0.2853	0.2693	0.3660	0.3127
	30	0.3753	0.4427	0.4627	0.3300	0.4220	0.5067	0.4300	0.3860	0.4780	0.4467
	50	0.6153	0.6327	0.6553	0.5633	0.6100	0.6713	0.6527	0.5900	0.6747	0.6293
	100	0.8860	0.8760	0.8853	0.8393	0.8713	0.8740	0.8933	0.8573	0.8880	0.8813
	200	0.9867	0.9827	0.9847	0.9800	0.9833	0.9820	0.9880	0.9827	0.9860	0.9873
0.12	10	0.0780	0.2693	0.2993	0.1947	0.2287	0.2273	0.1807	0.1913	0.3160	0.2500
	20	0.3347	0.4280	0.4527	0.3067	0.3947	0.3547	0.5113	0.3767	0.5713	0.5327
	30	0.5347	0.6160	0.6400	0.4527	0.5767	0.5627	0.5627	0.5580	0.6353	0.5747
	50	0.7467	0.7680	0.7873	0.6820	0.7493	0.7500	0.7620	0.7393	0.7927	0.7567
	100	0.9620	0.9607	0.9627	0.9293	0.9613	0.9573	0.9513	0.9560	0.9527	0.9453
	200	1.0000	0.9993	0.9993	0.9987	0.9993	0.9993	0.9993	0.9993	0.9993	0.9993

*Notes: t, t statistic; McKay, McKay; MiL, Miller; SK, Sharma and Krishna; CP, Curto and Pinto; NB, Non-parametric bootstrap; PB, Parametric bootstrap; BMiL1, Miller bootstrap 1; BMiL2, Miller bootstrap 2; BCP, Bootstrap Curto and Pinto

TESTING THE POPULATION COEFFICIENT OF VARIATION

Table 3.4: Estimated Power of Various Tests for the Log-Normal (2, 0.4724) Distribution

c	n	Tests for CV*									
		t	McKay	MiL	SK	CP	NB	PB	BMiL1	BMiL2	BCP
0.04	10	0.0220	0.0593	0.0740	0.0733	0.0733	0.0287	0.0333	0.0293	0.0707	0.0707
	20	0.0573	0.0867	0.0973	0.0873	0.0780	0.0653	0.0993	0.0473	0.1140	0.1027
	30	0.1100	0.1240	0.1400	0.1293	0.1013	0.1273	0.1540	0.0873	0.1613	0.1367
	50	0.1713	0.1660	0.1787	0.1853	0.1107	0.1707	0.1813	0.1253	0.1813	0.1180
	100	0.2647	0.2387	0.2480	0.2667	0.1600	0.2747	0.2720	0.2147	0.2513	0.1633
	200	0.4220	0.3680	0.3833	0.3987	0.2560	0.4287	0.4333	0.3427	0.3920	0.2707
0.06	10	0.0287	0.0747	0.0887	0.0780	0.0980	0.1067	0.1027	0.0347	0.1447	0.1607
	20	0.1140	0.1487	0.1620	0.1367	0.1380	0.1653	0.1393	0.1067	0.1667	0.1447
	30	0.1527	0.1687	0.1867	0.1540	0.1447	0.1447	0.1840	0.1300	0.2040	0.1567
	50	0.2347	0.2340	0.2513	0.2400	0.1813	0.2927	0.2300	0.1967	0.2353	0.1660
	100	0.4360	0.4060	0.4227	0.4147	0.3247	0.4860	0.4293	0.3840	0.4100	0.3120
	200	0.6607	0.6120	0.6253	0.6147	0.5120	0.6720	0.6147	0.5893	0.5787	0.4313
0.08	10	0.0513	0.1200	0.1360	0.1133	0.1347	0.0553	0.1107	0.0740	0.1640	0.1700
	20	0.1440	0.2013	0.2167	0.1560	0.1893	0.2013	0.1653	0.1407	0.2173	0.1893
	30	0.2273	0.2653	0.2853	0.2140	0.2393	0.2453	0.2667	0.2127	0.2960	0.2573
	50	0.3753	0.3813	0.4013	0.3600	0.3147	0.3560	0.4000	0.3313	0.4080	0.3333
	100	0.6253	0.6027	0.6153	0.5800	0.5047	0.6427	0.6560	0.5680	0.6347	0.5467
	200	0.8640	0.8327	0.8433	0.8173	0.7893	0.8607	0.8480	0.8207	0.8300	0.7633
0.10	10	0.0640	0.1533	0.1760	0.1273	0.1640	0.0953	0.1133	0.0967	0.1940	0.1773
	20	0.2147	0.2940	0.3087	0.2113	0.2773	0.2420	0.2400	0.2213	0.3080	0.2727
	30	0.3280	0.3813	0.3980	0.2920	0.3480	0.3500	0.3673	0.3193	0.4053	0.3587
	50	0.5053	0.5247	0.5400	0.4560	0.4487	0.4940	0.5167	0.4660	0.5380	0.4447
	100	0.7980	0.7860	0.7967	0.7420	0.7387	0.8100	0.8027	0.7627	0.7967	0.7400
	200	0.9673	0.9560	0.9600	0.9387	0.9413	0.9560	0.9573	0.9540	0.9527	0.9300
0.12	10	0.0693	0.1880	0.2093	0.1333	0.1960	0.1647	0.2560	0.1247	0.3213	0.3180
	20	0.2873	0.3787	0.3987	0.2533	0.3440	0.3473	0.4353	0.3100	0.4847	0.4533
	30	0.4213	0.4900	0.5053	0.3433	0.4400	0.4560	0.4307	0.4167	0.4973	0.4280
	50	0.6787	0.7067	0.7233	0.6100	0.6533	0.7353	0.6500	0.6580	0.6873	0.6113
	100	0.9100	0.9093	0.9133	0.8707	0.8807	0.9140	0.9253	0.8980	0.9227	0.8953
	200	0.9940	0.9940	0.9940	0.9900	0.9913	0.9940	0.9947	0.9940	0.9947	0.9940

*Notes: t, t statistic; McKay, McKay; MiL, Miller; SK, Sharma and Krishna; CP, Curto and Pinto; NB, Non-parametric bootstrap; PB, Parametric bootstrap; BMiL1, Miller bootstrap 1; BMiL2, Miller bootstrap 2; BCP, Bootstrap Curto and Pinto

- Creticos, P. S., Adams, W. R., & Petty, B. G. (2002). A methacholine challenge dose-response study for development of a pharmacodynamic bioequivalence methodology for albuterol metered-dose inhalers. *Journal of Allergy Clinical Immunization*, 100, 713-720.
- Curto, J. D., & Pinto, J. C. (2009a). The coefficient of variation asymptotic distribution in the case of non-iid random variables. *Journal of Applied Statistics*, 36(1), 21-32.
- Curto, J. D., Pinto, J. C., & Tavares, G. N. (2009b). Modeling stock market volatility using GARCH models with normal, student's t and stable Paretian distributions. *Statistical Papers*, 50, 311-321.
- Dodd, M. G., & Murphy, T. M. (1995). Accuracy and precision of techniques for counting great blue heron nests. *Journal of Wildlife Management*, 59, 667-673.
- Efron, B. (1979). Bootstrap methods: Another look at the jack knife. *Annals of Statistics*, 7, 1-26.
- Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. New York, NY: Chapman & Hall.
- Gong, J., & Li, Y. (1999). Relationship between estimated Weibull modulus and the coefficient of variation of the measured strength for ceramics. *Journal of the American Ceramic Society*, 82(2), 449-452.
- Hamer, A. J., Strachan, J. R., Black, M. M., Ibbotson, C., & Elson, R. A. (1995). A new method for comparative bone strength measurement. *Journal of Medical Engineering and Technology*, 19(1), 1-5.
- Hendricks, W. A., & Robey, W. K. (1936). The sampling distribution of the coefficient of variation. *Annals of Mathematical Statistics*, 1(7), 129-132.
- Iglewicz, B. (1967). Some properties of the coefficient of variation. Unpublished PhD thesis, Virginia Polytechnic Institute.
- Koopmans, L. H., Owen, D. B., & Rosenblatt, J. I. (1964). Confidence intervals for the coefficient of variation for the normal and lognormal distributions. *Biometrika*, 51, 25-32.
- McKay, A. T. (1932). Distribution of the coefficient of variation and the extended t distribution. *Journal of the Royal Statistical Society*, 95, 695-698.
- Miller, E. G. (1991). Asymptotic test statistics for coefficient of variation. *Communications in Statistics-Theory & Methods*, 20, 3351-3363.
- Miller, E. G., & Karson, M. J. (1977). Testing the equality of two coefficient of variation. *American Statistical Association: Proceedings of the Business and Economics Section, Part 1*, 278-283.
- Nairy, K. S., & Rao, K. A. (2003). Tests of coefficients of variation of normal population. *Communications in Statistics – Simulation and Computation*, 32, 641-661.
- Pearson, K. (1896). Mathematical contributions to the theory of evolution? III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society, Series A*, 187, 253-318.
- Pearson, E. S. (1932). Comparison of A.T. McKay's approximation with experimental sampling results. *Journal of the Royal Statistical Society*, 95, 703-704.
- Shi, W., & Kibria, B. M. G. (2007). On some confidence intervals for estimating the mean of a skewed population. *International Journal of Mathematical Education in Science and Technology*, 38, 412-421.
- Sharma, K. K., & Krishna, H. (1994). Asymptotic sampling distribution of inverse coefficient of variation and its applications. *IEEE Transaction Reliability*, 43, 630-633.

Class(es) of Factor-Type Estimator(s) in Presence of Measurement Error

Diwakar Shukla Sharad Pathak Narendra Singh Thakur
Dr. Hari Singh Gour University, Banasthali University,
Sagar, M. P., India Rajasthan, India

When data is collected via sample survey it is assumed whatever is reported by a respondent is correct. However, given the issues of prestige bias, personal respect and honor, respondents' self-reported data often produces over- or under- estimated values as opposed to true values regarding the variables under question. This causes measurement error to be present in sample values. This article considers the factor-type estimator as an estimation tool and examines its performance under a measurement error model. Expressions of optimization are derived and theoretical results are supported by numerical examples.

Key words: Measurement error, factor-type estimator, bias, mean squared error.

Introduction

Sample surveys result in an efficiency of estimators on the basis of collected or simulated data. Data for analyses may originate from various sampling sources, such as, simple random sampling, stratified sampling, systematic sampling or cluster sampling. Estimation methods are typically analyzed under the assumption that observations collected are true and without error; however, real life data, gathered through sample surveys contains errors due to memory failure, prestige bias, over reporting patterns, unwillingness to respond, desire for secrecy and other reasons. The deviation between true and observed values is error and is technically termed measurement error. Measurement error may be characterized as the difference between the value of a variable provided by the respondent and the true value of the same variable. The total survey error of a statistic with measurement error has both fixed error (bias) and variable error (variance) over repeated trials of the survey (Cochran, 2005;

Sukhatme, et al., 1984). Figure 1 illustrates the concept of measurement error.

There are two possibilities for incompleteness in a survey: incorrect response or non-response. Measurement bias provides a systematic pattern in the difference between the respondent's answers to a question and the correct answer. For example, a respondent may forget to report a few specific income sources resulting in total reported income being lower than actual. Measurement variance reflects random variation in answers provided to an interviewer while asking the same question, that is, often the same respondent provides different answers to the same question when asked repeatedly. Several methods are available in the survey sampling literature to handle non-response, including the revisit method, imputation methods, auxiliary sources utilization method and the neighboring units manipulation methods, however, when a respondent provides incorrect information regarding a variable, additional techniques are required. This study considers this aspect and deals with mean estimation under measurement error.

Manisha and Singh (2001) examined population mean estimation in the presence of measurement errors; they provided an effect of measurement errors on a new estimator obtained as a combination of ratio and mean per unit estimator. Shalabh (1997) studied a ratio method of estimation in the presence of measurement errors. Singh and Shukla (1987) presented a

Diwakar Shukla is an Associate Professor in the Department of Mathematics and Statistics. Email him at: diwakarshukla@rediffmail.com. Sharad Pathak is a Research Scholar in the Department of Mathematics and Statistics. Email him at: sharadpathakstats@yahoo.com. Narendra S. Thakur is an Assistant Professor at Banasthali University, Rajasthan.

family of factor-type ratio estimators. Shukla (2002) proposed a new strategy for estimation in the form of a factor-type ratio estimator in two phase sampling. Shukla, et al. (2009) also proposed a mean estimation under imputation of missing data using factor-type estimator in two-phase sampling and have since suggested a linear combination based imputation method for missing data in sample (Shukla, et al., 2011). Shukla, et al. (2012A) proposed an estimation of population mean using two auxiliary sources in sample surveys. Shukla, et al. (2012B) suggested an estimator for mean estimation in the presence measurement error of observations. Shukla, et al. (2012C) presented a transformed estimator for estimation of population mean with missing data in sample-surveys. Thakur, et al. (2011, 2012) suggested imputation strategies under double sampling.

Figure 1: Concept of Measurement Error

Singh and Karpe (2008a) presented a ratio-product estimator for population mean in the presence of measurement errors, Neter (1970) examined measurement errors in reports of consumer expenditures, Sud and Shrivastava (2000) studied estimation and population mean in repeat surveys on the presence of measurement errors and Sud, et al. (2001) considered a biased estimator in repeat surveys. Dalabehara and Sahoo (2000) and Kadilar and Cingi (2005) suggested estimators using two auxiliary sources in survey sampling. Other useful contributions over applications of measurement error models are provided by Fuller (1987), Cochran (1993), Mukhopadhyay

(2000), Murthy (1977), Sukhatme et al. (1984) and Cheng and Van Ness (1999). This article presents an estimation strategy under a measurement error model using two auxiliary sources for the purpose of optimization.

Study Notations and Assumptions

Assume a set of information obtained via a simple random sampling procedure on three characteristics Y , X_1 and X_2 . Suppose (y_i, x_{1i}, x_{2i}) are observational values and (Y_i, X_{1i}, X_{2i}) are corresponding true values for the characteristics respectively. Notations for this study are:

$\bar{Y}$, $\bar{X}_1$ and $\bar{X}_2$: Population parameters;

$\bar{y}$, $\bar{x}_1$ and $\bar{x}_2$: Mean per unit estimates for a simple random sample of size n ;

n : Sample size;

f : Sampling friction ($f = n/N$);

N : Population size;

U_i : Measurement error for Y ;

V_i : Measurement error for X_1 ;

T_i : Measurement error for X_2 ;

σ_U^2 , σ_V^2 and σ_T^2 : Variances for measurement error;

σ_Y^2 , $\sigma_{X_1}^2$ and $\sigma_{X_2}^2$: Variances of variable Y , X_1 and X_2 respectively;

ρ_{01} : Correlation between variable Y and X_1 ;

ρ_{02} : Correlation between variable Y and X_2 ;

ρ_{12} : Correlation between variable X_1 and X_2 ;

$C_Y = \sigma_Y / \bar{Y}$: Coefficient of variation for variable $Y(C_0)$;

$C_{X_1} = \sigma_{X_1} / \bar{X}_1$: Coefficient of variation for variable $X_1(C_1)$; and

$C_{X_2} = \sigma_{X_2} / \bar{X}_2$: Coefficient of variation for variable $X_2(C_2)$.

New notations are:

$$W_Y = \frac{\sum (Y_i - \bar{Y})}{\sqrt{n}},$$

$$W_{X_1} = \frac{\sum (X_{1i} - \bar{X}_1)}{\sqrt{n}},$$

$$W_{X_2} = \frac{\sum (X_{2i} - \bar{X}_2)}{\sqrt{n}},$$

$$W_U = \frac{\sum (U_i)}{\sqrt{n}},$$

$$W_V = \frac{\sum (V_i)}{\sqrt{n}},$$

and

$$W_T = \frac{\sum (T_i)}{\sqrt{n}}.$$

Assume the measurement errors are stochastic in nature and are uncorrelated, the sum of measurement error is zero and the variances are σ_U^2 , σ_V^2 and σ_T^2 , respectively. For an i^{th} unit ($i = 1, 2, 3, \dots, n$) unit in the sample assume the measurement errors are:

$$\left. \begin{aligned} U_i &= y_i - Y_i \\ V_i &= x_{1i} - X_{1i} \\ T_i &= x_{2i} - X_{2i} \end{aligned} \right\} \quad (3.1)$$

and, from (3.1),

$$\bar{y}_i = \frac{1}{n} \sum_i (U_i + Y_i) \quad (3.2)$$

or

$$\bar{y} - \bar{Y} = \frac{W_Y}{\sqrt{n}} + \frac{W_U}{\sqrt{n}}.$$

Similarly,

$$\bar{x}_1 - \bar{X}_1 = \frac{W_{X_1}}{\sqrt{n}} + \frac{W_V}{\sqrt{n}}$$

and

$$\bar{x}_2 - \bar{X}_2 = \frac{W_{X_2}}{\sqrt{n}} + \frac{W_T}{\sqrt{n}}.$$

Existing Estimators: Mean per Unit Estimator

The mean per unit (or mean) estimator is a well-known estimator, and in the setup of measurement error, $\bar{y}_i = n^{-1} \sum_i (U_i + Y_i)$, is shown in (3.2). The bias for $\bar{y}$ is zero, that is,

$$E(\bar{y}) = E\left[\frac{1}{n} \sum_i (U_i + Y_i)\right] = (\bar{Y}) \quad (4.1a)$$

and the variance is

$$Variance(\bar{y}) = \frac{\sigma_Y^2}{n} \left[1 + \frac{\sigma_U^2}{\sigma_Y^2} \right] \quad (4.1b)$$

To estimate $\bar{Y}$, the sample statistic $\bar{y}$, which provides an unbiased estimator, can be used. In mean per unit estimator $\bar{y}$ no additional information is required. Several methods exist for using the auxiliary X characteristic.

Existing Estimators: Shalabh (1997) Estimator

Shalabh (1997) proposed an estimator that is a ratio estimator studied under measurement error.

$$t_R = \frac{\bar{y}}{\bar{x}} \cdot \mu_X \quad (4.2)$$

Where the bias of t_R is

$$B(t_R) = \frac{\mu_Y}{n} \left[C_X (C_X - \rho C_Y) + \frac{\sigma_v^2}{\mu_X^2} \right] \quad (4.2a)$$

and the mean squared error is

$$\text{MSE}(t_R) = \frac{\sigma_Y^2}{n} \left[1 - \frac{C_X}{C_Y} \left(2\rho - \frac{C_X}{C_Y} \right) \right] + \frac{1}{n} \left[\sigma_u^2 + \left(\frac{\mu_Y}{\mu_X} \right)^2 \sigma_v^2 \right] \quad (4.2b)$$

where μ_X denotes the population mean of X .

Existing Estimators: Manisha and Singh (2001) Estimator

Manisha and Singh (2001) proposed the estimator

$$\bar{y}_\theta = \theta t_R + (1 - \theta) \bar{y} \quad (4.3)$$

where the bias of $\bar{y}_\theta$ is

$$B(\bar{y}_\theta) = \theta \left\{ \frac{\mu_Y}{n\mu_X^2} (\sigma_X^2 + \sigma_v^2) - \frac{1}{n\mu_X} \rho \sigma_X \sigma_Y \right\} \quad (4.3a)$$

and the mean squared error is

$$B(\bar{y}_\theta) = \frac{\sigma_Y^2}{n} \left[1 - \theta \frac{C_X}{C_Y} \left(2\rho - \theta \frac{C_X}{C_Y} \right) \right] + \frac{1}{n} \left[\theta^2 \frac{\mu_Y^2}{\mu_X^2} \sigma_v^2 + \sigma_u^2 \right] \quad (4.3b)$$

where θ is a characterizing scalar and U and V are measurement errors corresponding to Y and X respectively.

Proposed Estimator(s)

The two parameter F-T estimators proposed are:

$$\left. \begin{aligned} \bar{y}_{FT1}^\bullet &= \bar{y} T_1 T_2 \\ \bar{y}_{FT2}^\bullet &= \bar{y} T_1 T_2^{-1} \\ \bar{y}_{FT3}^\bullet &= \bar{y} T_1^{-1} T_2 \end{aligned} \right\} \quad (5.1)$$

where

$$f = n/N$$

and

$$\left. \begin{aligned} T_i &= \frac{(A_i + C_i) \bar{X}_i + f B_i \bar{X}_i}{(A_i + f B_i) \bar{X}_i + C_i \bar{X}_i} \\ A &= (K-1)(K-2) \\ B &= (K-1)(K-4) \\ C &= (K-2)(K-3)(K-4) \end{aligned} \right\} \quad (5.1b)$$

Thus,

$$\bar{y}_{FT1}^\bullet = \bar{y} \frac{(A_1 + C_1) \bar{X}_1 + f B_1 \bar{X}_1}{(A_1 + f B_1) \bar{X}_1 + C_1 \bar{X}_1} \frac{(A_2 + C_2) \bar{X}_2 + f B_2 \bar{X}_2}{(A_2 + f B_2) \bar{X}_2 + C_2 \bar{X}_2}, \quad (5.2a)$$

$$\bar{y}_{FT2}^\bullet = \bar{y} \frac{(A_1 + C_1) \bar{X}_1 + f B_1 \bar{X}_1}{(A_1 + f B_1) \bar{X}_1 + C_1 \bar{X}_1} \frac{(A_2 + f B_2) \bar{X}_2 + C_2 \bar{X}_2}{(A_2 + C_2) \bar{X}_2 + f B_2 \bar{X}_2} \quad (5.2b)$$

and

$$\bar{y}_{FT3}^\bullet = \bar{y} \frac{(A_1 + f B_1) \bar{X}_1 + C_1 \bar{X}_1}{(A_1 + C_1) \bar{X}_1 + f B_1 \bar{X}_1} \frac{(A_2 + C_2) \bar{X}_2 + f B_2 \bar{X}_2}{(A_2 + f B_2) \bar{X}_2 + C_2 \bar{X}_2} \quad (5.2c)$$

Table 5.1: Members of the Proposed Class(es)

K_1	K_2			
	1	2	3	4
1	$t_1 = \bar{y} \frac{\bar{X}_1}{\bar{x}_1} \frac{\bar{X}_2}{\bar{x}_2}$	$t_2 = \bar{y} \frac{\bar{X}_1}{\bar{x}_1} \frac{\bar{x}_2}{\bar{X}_2}$	$t_3 = \bar{y} \frac{\bar{X}_1}{\bar{x}_1} \frac{N\bar{X}_2 - n\bar{x}_2}{(N-n)\bar{X}_2}$	$t_4 = \bar{y} \frac{\bar{X}_1}{\bar{x}_1}$
2	$t_5 = \bar{y} \frac{\bar{x}_1}{\bar{X}_1} \frac{\bar{X}_2}{\bar{x}_2}$	$t_6 = \bar{y} \frac{\bar{x}_1}{\bar{X}_1} \frac{\bar{x}_2}{\bar{X}_2}$	$t_7 = \bar{y} \frac{\bar{x}_1}{\bar{X}_1} \frac{N\bar{X}_2 - n\bar{x}_2}{(N-n)\bar{X}_2}$	$t_8 = \bar{y} \frac{\bar{x}_1}{\bar{X}_1}$
3	$t_9 = \bar{y} \frac{N\bar{X}_1 - n\bar{x}_1}{(N-n)\bar{X}_1} \frac{\bar{X}_2}{\bar{x}_2}$	$t_{10} = \bar{y} \frac{N\bar{X}_1 - n\bar{x}_1}{(N-n)\bar{X}_1} \frac{\bar{x}_2}{\bar{X}_2}$	$t_{11} = \bar{y} \frac{N\bar{X}_1 - n\bar{x}_1}{(N-n)\bar{X}_1} \frac{N\bar{X}_2 - n\bar{x}_2}{(N-n)\bar{X}_2}$	$t_{12} = \bar{y} \frac{N\bar{X}_1 - n\bar{x}_1}{(N-n)\bar{X}_1}$
4	$t_{13} = \bar{y} \frac{\bar{X}_2}{\bar{x}_2}$	$t_{14} = \bar{y} \frac{\bar{x}_2}{\bar{X}_2}$	$t_{15} = \bar{y} \frac{N\bar{X}_2 - n\bar{x}_2}{(N-n)\bar{X}_2}$	$\bar{y}$

Note that there is a combination of K_i where $i = (1, 2)$ where ($K_1 = K_2$) (see Table 5.1 for factors). When K_i where $i = (1, 2)$ is constant, it is important to choose suitably so that the resultant mean squared error of the proposed estimators may be minimized to the greatest extent. Using the proposed estimator many different estimators may be obtained because an estimator exists for each combination of (K_1, K_2).

Properties of the Proposed Estimator(s)

For the approximation assume that:

$$\delta_0 = \frac{1}{\sqrt{n}} (W_Y + W_U);$$

$$\delta_1 = \frac{1}{\sqrt{n}} (W_{X_1} + W_V);$$

$$\delta_2 = \frac{1}{\sqrt{n}} (W_{X_2} + W_T);$$

$$\theta_i = \alpha_i - \beta_i;$$

$$f = \frac{n}{N};$$

$$\alpha_i = \frac{fB_i}{A_i + fB_i + C_i};$$

and

$$\beta_i = \frac{C_i}{A_i + fB_i + C_i}.$$

Theorem 6.1

The estimator $\bar{y}_{FT1}^\bullet$ up to first order of approximation can be expressed as:

$$\begin{aligned} \bar{y}_{FT1}^\bullet &= \bar{Y} + \delta_0 + \frac{\bar{Y}\theta_1}{\bar{X}_1} \delta_1 + \frac{\bar{Y}\theta_2}{\bar{X}_2} \delta_2 - \frac{\bar{Y}\beta_1\theta_1}{\bar{X}_1^2} \delta_1^2 \\ &\quad - \frac{\bar{Y}\beta_2\theta_2}{\bar{X}_2^2} \delta_2^2 + \frac{\theta_1}{\bar{X}_1} \delta_0 \delta_1 + \frac{\theta_2}{\bar{X}_2} \delta_0 \delta_2 + \frac{\bar{Y}\theta_1\theta_2}{\bar{X}_1\bar{X}_2} \delta_1 \delta_2 \end{aligned} \quad (6.1)$$

and the bias of $\bar{y}_{FT1}^\bullet$ is:

$$\begin{aligned} \text{Bias}(\bar{y}_{FT1}^{\bullet}) &= \frac{\bar{Y}}{n} \left\{ \theta_1 \rho_{01} C_0 C_1 + \theta_2 \rho_{02} C_0 C_2 \right. \\ &\quad \left. + \theta_1 \theta_2 \rho_{12} C_1 C_2 - \beta_1 \theta_1 C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) - \beta_2 \theta_2 C_2^2 \left(1 + \frac{\sigma_r^2}{\sigma_{x_2}^2} \right) \right\} \end{aligned} \quad (6.2)$$

The mean squared error of $\bar{y}_{FT1}^{\bullet}$ is:

$$\begin{aligned} \text{MSE}(\bar{y}_{FT1}^{\bullet}) &= \\ &= \frac{1}{n} (\sigma_y^2 + \sigma_u^2) + \frac{\bar{Y}^2}{n} \left[\theta_1^2 C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) \right. \\ &\quad \left. + \theta_2^2 C_2^2 \left(1 + \frac{\sigma_r^2}{\sigma_{x_2}^2} \right) + 2\theta_1 \rho_{01} C_0 C_1 \right. \\ &\quad \left. + 2\theta_2 \rho_{02} C_0 C_2 + 2\theta_1 \theta_2 \rho_{12} C_1 C_2 \right] \end{aligned} \quad (6.3)$$

Proof 6.1

From (5.2a) the proposed estimator is

$$\bar{y}_{FT1}^{\bullet} = \frac{\bar{y} (A_1 + C_1) \bar{X}_1 + f B_1 \bar{x}_1 (A_2 + C_2) \bar{X}_2 + f B_2 \bar{x}_2}{(A_1 + f B_1) \bar{X}_1 + C_1 \bar{x}_1 (A_2 + f B_2) \bar{X}_2 + C_2 \bar{x}_2}$$

$$\begin{aligned} \bar{y}_{FT1}^{\bullet} &= [\bar{Y} + \delta_0] \left\{ 1 + \frac{\alpha_1 \delta_1}{\bar{X}_1} \right\} \left\{ 1 + \frac{\alpha_2 \delta_2}{\bar{X}_2} \right\} \\ &\quad \left\{ 1 + \frac{\beta_1 \delta_1}{\bar{X}_1} \right\}^{-1} \left\{ 1 + \frac{\beta_2 \delta_2}{\bar{X}_2} \right\}^{-1} \end{aligned}$$

and from this,

$$\begin{aligned} \bar{y}_{FT1}^{\bullet} &= \bar{Y} + \delta_0 + \frac{\bar{Y} \theta_1}{\bar{X}_1} \delta_1 + \frac{\bar{Y} \theta_2}{\bar{X}_2} \delta_2 \\ &\quad - \frac{\bar{Y} \beta_1 \theta_1}{\bar{X}_1^2} \delta_1^2 - \frac{\bar{Y} \beta_2 \theta_2}{\bar{X}_2^2} \delta_2^2 \\ &\quad + \frac{\theta_1}{\bar{X}_1} \delta_0 \delta_1 + \frac{\theta_2}{\bar{X}_2} \delta_0 \delta_2 + \frac{\bar{Y} \theta_1 \theta_2}{\bar{X}_1 \bar{X}_2} \delta_1 \delta_2 \end{aligned}$$

and

$$\begin{aligned} \bar{y}_{FT1}^{\bullet} - \bar{Y} &= \delta_0 + \frac{\bar{Y} \theta_1}{\bar{X}_1} \delta_1 + \frac{\bar{Y} \theta_2}{\bar{X}_2} \delta_2 \\ &\quad - \frac{\bar{Y} \beta_1 \theta_1}{\bar{X}_1^2} \delta_1^2 - \frac{\bar{Y} \beta_2 \theta_2}{\bar{X}_2^2} \delta_2^2 \\ &\quad + \frac{\theta_1}{\bar{X}_1} \delta_0 \delta_1 + \frac{\theta_2}{\bar{X}_2} \delta_0 \delta_2 + \frac{\bar{Y} \theta_1 \theta_2}{\bar{X}_1 \bar{X}_2} \delta_1 \delta_2. \end{aligned}$$

Thus, for the solution:

$$\begin{aligned} [\bar{y}_{FT1}^{\bullet} - \bar{Y}]^2 &= \left\{ \delta_0 + \frac{\bar{Y} \theta_1}{\bar{X}_1} \delta_1 + \frac{\bar{Y} \theta_2}{\bar{X}_2} \delta_2 \right. \\ &\quad - \frac{\bar{Y} \beta_1 \theta_1}{\bar{X}_1^2} \delta_1^2 - \frac{\bar{Y} \beta_2 \theta_2}{\bar{X}_2^2} \delta_2^2 + \frac{\theta_1}{\bar{X}_1} \delta_0 \delta_1 \\ &\quad \left. + \frac{\theta_2}{\bar{X}_2} \delta_0 \delta_2 + \frac{\bar{Y} \theta_1 \theta_2}{\bar{X}_1 \bar{X}_2} \delta_1 \delta_2 \right\}^2 \end{aligned}$$

and

$$\begin{aligned} E[\bar{y}_{FT1}^{\bullet} - \bar{Y}]^2 &= \frac{1}{n} (\sigma_y^2 + \sigma_u^2) \\ &\quad + \frac{\bar{Y}^2}{n} \left[\theta_1^2 C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) + \theta_2^2 C_2^2 \left(1 + \frac{\sigma_r^2}{\sigma_{x_2}^2} \right) \right. \\ &\quad \left. + 2\theta_1 \rho_{01} C_0 C_1 + 2\theta_2 \rho_{02} C_0 C_2 + 2\theta_1 \theta_2 \rho_{12} C_1 C_2 \right]. \end{aligned}$$

Theorem 6.2

The estimator $\bar{y}_{FT2}^{\bullet}$ up to first order of approximation can be expressed as:

$$\begin{aligned} \bar{y}_{FT2}^{\bullet} &= \bar{Y} + \delta_0 + \frac{\bar{Y} \theta_1}{\bar{X}_1} \delta_1 - \frac{\bar{Y} \theta_2}{\bar{X}_2} \delta_2 - \frac{\bar{Y} \beta_1 \theta_1}{\bar{X}_1^2} \delta_1^2 \\ &\quad + \frac{\bar{Y} \alpha_2 \theta_2}{\bar{X}_2^2} \delta_2^2 + \frac{\theta_1}{\bar{X}_1} \delta_0 \delta_1 - \frac{\theta_2}{\bar{X}_2} \delta_0 \delta_2 - \frac{\bar{Y} \theta_1 \theta_2}{\bar{X}_1 \bar{X}_2} \delta_1 \delta_2. \end{aligned} \quad (6.4)$$

The bias of $\bar{y}_{FT2}^{\bullet}$ is:

$$\begin{aligned}
 \text{Bias}(\bar{y}_{FT2}^\bullet) = & \frac{\bar{Y}}{n} \left\{ \theta_1 \rho_{01} C_0 C_1 - \theta_2 \rho_{02} C_0 C_2 - \theta_1 \theta_2 \rho_{12} C_1 C_2 \right. \\
 & \left. - \beta_1 \theta_1 C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) + \alpha_2 \theta_2 C_2^2 \left(1 + \frac{\sigma_T^2}{\sigma_{x_2}^2} \right) \right\} \quad (6.5)
 \end{aligned}$$

and the mean squared error of $\bar{y}_{FT2}^\bullet$ is:

$$\begin{aligned}
 \text{MSE}(\bar{y}_{FT2}^\bullet) = & \frac{1}{n} (\sigma_y^2 + \sigma_u^2) \\
 & + \frac{\bar{Y}^2}{n} \left[\theta_1^2 C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) + \theta_2^2 C_2^2 \left(1 + \frac{\sigma_T^2}{\sigma_{x_2}^2} \right) \right. \\
 & \left. + 2\theta_1 \rho_{01} C_0 C_1 - 2\theta_2 \rho_{02} C_0 C_2 - 2\theta_1 \theta_2 \rho_{12} C_1 C_2 \right]. \quad (6.6)
 \end{aligned}$$

Proof 6.2

From (5.2b) the proposed estimator is

$$\begin{aligned}
 \bar{y}_{FT2}^\bullet = & \frac{\bar{y} (A_1 + C_1) \bar{X}_1 + f B_1 \bar{x}_1 (A_2 + f B_2) \bar{X}_2 + C_2 \bar{x}_2}{(A_1 + f B_1) \bar{X}_1 + C_1 \bar{x}_1 (A_2 + C_2) \bar{X}_2 + f B_2 \bar{x}_2} \\
 \bar{y}_{FT2}^\bullet = & [\bar{Y} + \delta_0] \left\{ 1 + \frac{\alpha_1 \delta_1}{\bar{X}_1} \right\} \left\{ 1 + \frac{\beta_2 \delta_2}{\bar{X}_2} \right\} \\
 & \left\{ 1 + \frac{\alpha_2 \delta_2}{\bar{X}_2} \right\}^{-1} \left\{ 1 + \frac{\beta_1 \delta_1}{\bar{X}_1} \right\}^{-1}
 \end{aligned}$$

Solving the equations results in:

$$\begin{aligned}
 \bar{y}_{FT2}^\bullet = & \bar{Y} + \delta_0 + \frac{\bar{Y} \theta_1}{\bar{X}_1} \delta_1 - \frac{\bar{Y} \theta_2}{\bar{X}_2} \delta_2 \\
 & - \frac{\bar{Y} \beta_1 \theta_1}{\bar{X}_1^2} \delta_1^2 + \frac{\bar{Y} \alpha_2 \theta_2}{\bar{X}_2^2} \delta_2^2 \\
 & + \frac{\theta_1}{\bar{X}_1} \delta_0 \delta_1 - \frac{\theta_2}{\bar{X}_2} \delta_0 \delta_2 - \frac{\bar{Y} \theta_1 \theta_2}{\bar{X}_1 \bar{X}_2} \delta_1 \delta_2,
 \end{aligned}$$

$$\begin{aligned}
 \bar{y}_{FT2}^\bullet - \bar{Y} = & \delta_0 + \frac{\bar{Y} \theta_1}{\bar{X}_1} \delta_1 - \frac{\bar{Y} \theta_2}{\bar{X}_2} \delta_2 \\
 & - \frac{\bar{Y} \beta_1 \theta_1}{\bar{X}_1^2} \delta_1^2 + \frac{\bar{Y} \alpha_2 \theta_2}{\bar{X}_2^2} \delta_2^2 \\
 & + \frac{\theta_1}{\bar{X}_1} \delta_0 \delta_1 - \frac{\theta_2}{\bar{X}_2} \delta_0 \delta_2 - \frac{\bar{Y} \theta_1 \theta_2}{\bar{X}_1 \bar{X}_2} \delta_1 \delta_2,
 \end{aligned}$$

$$\begin{aligned}
 E[\bar{y}_{FT2}^\bullet - \bar{Y}] = & E \left[\delta_0 + \frac{\bar{Y} \theta_1}{\bar{X}_1} \delta_1 - \frac{\bar{Y} \theta_2}{\bar{X}_2} \delta_2 \right. \\
 & \left. - \frac{\bar{Y} \beta_1 \theta_1}{\bar{X}_1^2} \delta_1^2 + \frac{\bar{Y} \alpha_2 \theta_2}{\bar{X}_2^2} \delta_2^2 + \frac{\theta_1}{\bar{X}_1} \delta_0 \delta_1 \right. \\
 & \left. - \frac{\theta_2}{\bar{X}_2} \delta_0 \delta_2 - \frac{\bar{Y} \theta_1 \theta_2}{\bar{X}_1 \bar{X}_2} \delta_1 \delta_2 \right],
 \end{aligned}$$

$$\begin{aligned}
 E[\bar{y}_{FT2}^\bullet - \bar{Y}] = & \frac{\bar{Y}}{n} \left\{ \theta_1 \rho_{01} C_0 C_1 - \theta_2 \rho_{02} C_0 C_2 \right. \\
 & \left. - \theta_1 \theta_2 \rho_{12} C_1 C_2 - \beta_1 \theta_1 C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) \right. \\
 & \left. + \alpha_2 \theta_2 C_2^2 \left(1 + \frac{\sigma_T^2}{\sigma_{x_2}^2} \right) \right\}
 \end{aligned}$$

and

$$\begin{aligned}
 E[\bar{y}_{FT2}^\bullet - \bar{Y}]^2 = & E \left[\delta_0 + \frac{\bar{Y} \theta_1}{\bar{X}_1} \delta_1 - \frac{\bar{Y} \theta_2}{\bar{X}_2} \delta_2 \right. \\
 & \left. - \frac{\bar{Y} \beta_1 \theta_1}{\bar{X}_1^2} \delta_1^2 + \frac{\bar{Y} \alpha_2 \theta_2}{\bar{X}_2^2} \delta_2^2 + \frac{\theta_1}{\bar{X}_1} \delta_0 \delta_1 \right. \\
 & \left. - \frac{\theta_2}{\bar{X}_2} \delta_0 \delta_2 - \frac{\bar{Y} \theta_1 \theta_2}{\bar{X}_1 \bar{X}_2} \delta_1 \delta_2 \right]^2.
 \end{aligned}$$

Based on the solution:

$$\begin{aligned}
 E[\bar{y}_{FT2}^\bullet - \bar{Y}]^2 = & \frac{1}{n} (\sigma_y^2 + \sigma_u^2) \\
 & + \frac{\bar{Y}^2}{n} \left[\theta_1^2 C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) + \theta_2^2 C_2^2 \left(1 + \frac{\sigma_T^2}{\sigma_{x_2}^2} \right) \right. \\
 & \left. + 2\theta_1 \rho_{01} C_0 C_1 - 2\theta_2 \rho_{02} C_0 C_2 - 2\theta_1 \theta_2 \rho_{12} C_1 C_2 \right].
 \end{aligned}$$

Theorem 6.3

The estimator $\bar{y}_{FT3}^\bullet$, up to first order of approximation, can be expressed as:

$$\begin{aligned}\bar{y}_{FT3}^\bullet &= \bar{Y} + \delta_0 - \frac{\bar{Y}\theta_1}{\bar{X}_1}\delta_1 + \frac{\bar{Y}\theta_2}{\bar{X}_2}\delta_2 \\ &\quad + \frac{\bar{Y}\alpha_1\theta_1}{\bar{X}_1^2}\delta_1^2 - \frac{\bar{Y}\beta_2\theta_2}{\bar{X}_2^2}\delta_2^2 \\ &\quad - \frac{\theta_1}{\bar{X}_1}\delta_0\delta_1 + \frac{\theta_2}{\bar{X}_2}\delta_0\delta_2 - \frac{\bar{Y}\theta_1\theta_2}{\bar{X}_1\bar{X}_2}\delta_1\delta_2\end{aligned}\quad (6.7)$$

and the bias of $\bar{y}_{FT3}^\bullet$ is:

$$\begin{aligned}\text{Bias}(\bar{y}_{FT3}^\bullet) &= \frac{\bar{Y}}{n} \left\{ \theta_2\rho_{02}C_0C_2 - \theta_1\rho_{01}C_0C_1 \right. \\ &\quad - \theta_1\theta_2\rho_{12}C_1C_2 + \alpha_1\theta_1C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) \\ &\quad \left. - \beta_2\theta_2C_2^2 \left(1 + \frac{\sigma_t^2}{\sigma_{x_2}^2} \right) \right\}.\end{aligned}\quad (6.8)$$

The mean squared error of $\bar{y}_{FT3}^\bullet$ is:

$$\begin{aligned}\text{MSE}(\bar{y}_{FT3}^\bullet) &= \frac{1}{n}(\sigma_y^2 + \sigma_u^2) \\ &\quad + \frac{\bar{Y}^2}{n} \left[\theta_1^2C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) + \theta_2^2C_2^2 \left(1 + \frac{\sigma_t^2}{\sigma_{x_2}^2} \right) \right. \\ &\quad \left. - 2\theta_1\rho_{01}C_0C_1 + 2\theta_2\rho_{02}C_0C_2 - 2\theta_1\theta_2\rho_{12}C_1C_2 \right].\end{aligned}\quad (6.9)$$

Proof 6.3

From (5.2c) the proposed estimator is

$$\bar{y}_{FT3}^\bullet = \frac{\bar{y}(A_1 + fB_1)\bar{X}_1 + C_1\bar{x}_1(A_2 + C_2)\bar{X}_2 + fB_2\bar{x}_2}{(A_1 + C_1)\bar{X}_1 + fB_1\bar{x}_1(A_2 + fB_2)\bar{X}_2 + C_2\bar{x}_2}$$

$$\begin{aligned}\bar{y}_{FT3}^\bullet &= [\bar{Y} + \delta_0] \left\{ 1 + \frac{\alpha_2\delta_2}{\bar{X}_2} \right\} \left\{ 1 + \frac{\beta_1\delta_1}{\bar{X}_1} \right\} \\ &\quad \left\{ 1 + \frac{\alpha_1\delta_1}{\bar{X}_1} \right\}^{-1} \left\{ 1 + \frac{\beta_2\delta_2}{\bar{X}_2} \right\}^{-1},\end{aligned}$$

which results in

$$\begin{aligned}\bar{y}_{FT3}^\bullet &= \bar{Y} + \delta_0 - \frac{\bar{Y}\theta_1}{\bar{X}_1}\delta_1 + \frac{\bar{Y}\theta_2}{\bar{X}_2}\delta_2 \\ &\quad + \frac{\bar{Y}\alpha_1\theta_1}{\bar{X}_1^2}\delta_1^2 - \frac{\bar{Y}\beta_2\theta_2}{\bar{X}_2^2}\delta_2^2 \\ &\quad - \frac{\theta_1}{\bar{X}_1}\delta_0\delta_1 + \frac{\theta_2}{\bar{X}_2}\delta_0\delta_2 - \frac{\bar{Y}\theta_1\theta_2}{\bar{X}_1\bar{X}_2}\delta_1\delta_2,\end{aligned}$$

$$\begin{aligned}\bar{y}_{FT3}^\bullet - \bar{Y} &= \delta_0 - \frac{\bar{Y}\theta_1}{\bar{X}_1}\delta_1 + \frac{\bar{Y}\theta_2}{\bar{X}_2}\delta_2 \\ &\quad + \frac{\bar{Y}\alpha_1\theta_1}{\bar{X}_1^2}\delta_1^2 - \frac{\bar{Y}\beta_2\theta_2}{\bar{X}_2^2}\delta_2^2 \\ &\quad - \frac{\theta_1}{\bar{X}_1}\delta_0\delta_1 + \frac{\theta_2}{\bar{X}_2}\delta_0\delta_2 - \frac{\bar{Y}\theta_1\theta_2}{\bar{X}_1\bar{X}_2}\delta_1\delta_2,\end{aligned}$$

$$\begin{aligned}E[\bar{y}_{FT3}^\bullet - \bar{Y}] &= \\ E \left[\begin{aligned} &\delta_0 - \frac{\bar{Y}\theta_1}{\bar{X}_1}\delta_1 + \frac{\bar{Y}\theta_2}{\bar{X}_2}\delta_2 \\ &+ \frac{\bar{Y}\alpha_1\theta_1}{\bar{X}_1^2}\delta_1^2 - \frac{\bar{Y}\beta_2\theta_2}{\bar{X}_2^2}\delta_2^2 \\ &- \frac{\theta_1}{\bar{X}_1}\delta_0\delta_1 + \frac{\theta_2}{\bar{X}_2}\delta_0\delta_2 - \frac{\bar{Y}\theta_1\theta_2}{\bar{X}_1\bar{X}_2}\delta_1\delta_2 \end{aligned} \right]\end{aligned}$$

and

$$E[\bar{y}_{FT3}^{\bullet} - \bar{Y}] = \frac{\bar{Y}}{n} \left\{ \theta_2 \rho_{02} C_0 C_2 - \theta_1 \rho_{01} C_0 C_1 \right. \\ \left. - \theta_1 \theta_2 \rho_{12} C_1 C_2 + \alpha_1 \theta_1 C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) \right. \\ \left. - \beta_2 \theta_2 C_2^2 \left(1 + \frac{\sigma_T^2}{\sigma_{x_2}^2} \right) \right\}.$$

From which follows

$$[\bar{y}_{FT3}^{\bullet} - \bar{Y}]^2 = \left[\delta_0 - \frac{\bar{Y}\theta_1}{\bar{X}_1} \delta_1 + \frac{\bar{Y}\theta_2}{\bar{X}_2} \delta_2 + \frac{\bar{Y}\alpha_1\theta_1}{\bar{X}_1^2} \delta_1^2 \right. \\ \left. - \frac{\bar{Y}\beta_2\theta_2}{\bar{X}_2^2} \delta_2^2 - \frac{\theta_1}{\bar{X}_1} \delta_0 \delta_1 + \frac{\theta_2}{\bar{X}_2} \delta_0 \delta_2 - \frac{\bar{Y}\theta_1\theta_2}{\bar{X}_1\bar{X}_2} \delta_1 \delta_2 \right]^2$$

and

$$E[\bar{y}_{FT3}^{\bullet} - \bar{Y}]^2 = E \left[\delta_0 - \frac{\bar{Y}\theta_1}{\bar{X}_1} \delta_1 + \frac{\bar{Y}\theta_2}{\bar{X}_2} \delta_2 \right. \\ \left. + \frac{\bar{Y}\alpha_1\theta_1}{\bar{X}_1^2} \delta_1^2 - \frac{\bar{Y}\beta_2\theta_2}{\bar{X}_2^2} \delta_2^2 - \frac{\theta_1}{\bar{X}_1} \delta_0 \delta_1 \right. \\ \left. + \frac{\theta_2}{\bar{X}_2} \delta_0 \delta_2 - \frac{\bar{Y}\theta_1\theta_2}{\bar{X}_1\bar{X}_2} \delta_1 \delta_2 \right]^2.$$

The solution of which results in:

$$E[\bar{y}_{FT3}^{\bullet} - \bar{Y}]^2 = \frac{1}{n} (\sigma_v^2 + \sigma_u^2) \\ + \frac{\bar{Y}^2}{n} \left[\theta_1^2 C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) + \theta_2^2 C_2^2 \left(1 + \frac{\sigma_T^2}{\sigma_{x_2}^2} \right) \right. \\ \left. - 2\theta_1 \rho_{01} C_0 C_1 + 2\theta_2 \rho_{02} C_0 C_2 - 2\theta_1 \theta_2 \rho_{12} C_1 C_2 \right].$$

Minimum Mean Squared Error & Optimal Choices for the Proposed Estimator(s)

The mean squared error of the proposed estimators $\bar{y}_{FT1}^{\bullet}$, $\bar{y}_{FT2}^{\bullet}$ and $\bar{y}_{FT3}^{\bullet}$ shown in (6.3), (6.6) and (6.9) respectively, are functions with unknown parameter θ_i ; $i = (1, 2)$, whereas θ_i is a function of K solely. Thus, it is

practical to calculate an optimum value of K in such a way that the mean squared error of the resultant proposed estimator becomes least.

Consider $\bar{y}_{FT1}^{\bullet}$, notice the minimum mean squared error. On differentiation of $MSE(\bar{y}_{FT1}^{\bullet})$ with respect to θ_1 and θ_2 and equating to zero (assuming $\theta_i \neq 0$), two simultaneous equations result:

$$C_1^2 \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) \theta_1 + \rho_{12} C_1 C_2 \theta_2 + \rho_{01} C_0 C_1 = 0 \\ \text{for } \frac{\partial}{\partial \theta_1} [E[\bar{y}_{FT1}^{\bullet} - \bar{Y}]^2] = 0 \quad (7.1)$$

and

$$C_2^2 \left(1 + \frac{\sigma_T^2}{\sigma_{x_2}^2} \right) \theta_2 + \rho_{12} C_1 C_2 \theta_1 + \rho_{02} C_0 C_2 = 0 \\ \text{For } \frac{\partial}{\partial \theta_2} [E[\bar{y}_{FT1}^{\bullet} - \bar{Y}]^2] = 0. \quad (7.2)$$

From (7.1) and (7.2) the values of θ_1 and θ_2 are:

$$\left. \begin{aligned} \hat{\theta}_{(1)} &= \frac{C_0}{C_1} \frac{\rho_{02}\rho_{12} - \rho_{01} \left(1 + \frac{\sigma_T^2}{\sigma_{x_2}^2} \right)}{\left[\left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) \left(1 + \frac{\sigma_T^2}{\sigma_{x_2}^2} \right) - \rho_{12}^2 \right]} \\ \hat{\theta}_{(2)} &= \frac{C_0}{C_2} \frac{\rho_{01}\rho_{12} - \rho_{02} \left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right)}{\left[\left(1 + \frac{\sigma_v^2}{\sigma_{x_1}^2} \right) \left(1 + \frac{\sigma_T^2}{\sigma_{x_2}^2} \right) - \rho_{12}^2 \right]} \end{aligned} \right\} \quad (7.3)$$

The optimum values $\hat{\theta}_{(1)} = \Delta_1$ and $\hat{\theta}_{(2)} = \Delta_2$, for example, provide a minimum

mean squared error to $\bar{y}_{FT1}^\bullet$, where the second derivative is positive. Similarly, $\hat{\theta}_{(3)} = \hat{\theta}_{(1)}$; $\hat{\theta}_{(4)} = (-1)\hat{\theta}_{(2)}$ and $\hat{\theta}_{(5)} = (-1)\hat{\theta}_{(1)}$; $\hat{\theta}_{(6)} = \hat{\theta}_{(2)}$ are optimal choices corresponding to $\bar{y}_{FT2}^\bullet$ and $\bar{y}_{FT3}^\bullet$ respectively. These $\theta_{(\bullet)}$ provide polynomials in terms of K to produce values for which the mean squared error will be optimum.

Empirical Study

This illustration demonstrates how to evaluate the gain in efficiencies (in terms of mean squared error) obtained by the proposed estimators. To evaluate the performance of the various estimators discussed, a population is considered (see Appendix A); required information is shown in Table 8.1.

Table 8.1: Population Parameters

Parameter	Value	Parameter	Value
$\bar{Y}$	63.396	n	50
$\bar{X}_1$	48.136	N	250
$\bar{X}_2$	56.364	f	0.2
C_0	0.2899	ρ_{01}	0.8544
C_1	0.4637	ρ_{02}	0.8249
C_2	0.4085	ρ_{12}	0.8289

Table 8.2: Percent Relative Efficiency of Various Estimators with respect to Mean per Unit Estimator

Estimator(s)	$PRE(\bullet)$ with respect to MPU Estimator		
	$MSE(\bar{y}_{FT1}^\bullet)$	$MSE(\bar{y}_{FT2}^\bullet)$	$MSE(\bar{y}_{FT3}^\bullet)$
$\bar{y}$	100	100	100
t_1	40.75	27.41	42.31
t_2	27.41	40.75	23.63
t_3	93.26	64.96	109.88
t_4	36.30	36.29	21.95
t_5	21.35	11.52	18.71
t_6	11.53	21.35	22.65
t_7	23.87	19.27	33.28
t_8	21.95	21.95	36.29
t_9	38.35	28.38	40.94
t_{10}	28.38	38.35	21.79
t_{11}	104.25	84.15	79.65
t_{12}	106.28	106.2	71.89
t_{13}	42.40	25.70	42.39
t_{14}	25.71	42.39	25.70
t_{15}	106.23	75.51	106.21
Opt $(\bar{y}_{FT1}^\bullet)^*$	113.05	92.27	95.01
Opt $(\bar{y}_{FT2}^\bullet)^*$	92.28	113.04	74.28
Opt $(\bar{y}_{FT3}^\bullet)^*$	95.02	74.28	113.04

Results

Three different approaches were examined as tools for estimating in the presence of measurement error. Results indicate that the proposed approaches are effective and efficient over many existing strategies. The multiple choices for K are accessible via:

$$\begin{aligned} &(\Delta_1 + 1)K_1^3 + (f\Delta_1 - f - 8\Delta_1 - 9)K_1^2 \\ &+ (23\Delta_1 - 5f\Delta_1 + 5f + 26)K_1 \\ &+ (4f\Delta_1 - 22\Delta_1 - 4f - 24) = 0 \end{aligned} \quad (9.1)$$

and

$$\begin{aligned} &(\Delta_2 + 1)K_2^3 + (f\Delta_2 - f - 8\Delta_2 - 9)K_2^2 \\ &+ (23\Delta_2 - 5f\Delta_2 + 5f + 26)K_2 \\ &+ (4f\Delta_2 - 22\Delta_2 - 4f - 24) = 0. \end{aligned} \quad (9.2)$$

Polynomials (9.1) and (9.2), which are obtained from (7.3), provide three roots for lesser mean squared error. As discussed for $K_1 = K_2 = 4$ the proposed classes provide mean per unit estimators, thus those values are unbiased estimators.

For the estimator $\bar{y}_{FT1}^\bullet$, the optimum values of the characterizing scalar are $(K_1)_1 = 4.4951$, $(K_1)_2 = 3.1167$, $(K_1)_3 = 1.8111$, $(K_2)_1 = 4.5063$, $(K_2)_2 = 3.1133$ and $(K_2)_3 = 1.8096$. For $\bar{y}_{FT2}^\bullet$, the values are $(K_1)_4 = (K_1)_1$, $(K_1)_5 = (K_1)_2$, $(K_1)_6 = (K_1)_3$ and $(K_2)_4 = 1.8857$, and for the $\bar{y}_{FT3}^\bullet$ estimator values are, $(K_1)_7 = 11.5039$, $(K_2)_7 = (K_2)_1$, $(K_2)_8 = (K_2)_2$ and $(K_2)_9 = (K_2)_3$ with the remaining imaginary roots.

Tables 8.2 and 8.3 show that the proposed estimator is efficient over many currently used estimators, including the Manisha and Singh (2001) and the Shalabh (1997).

Conclusion

Based on study results, the proposed estimator(s) have several benefits over estimators currently used in research, including:

1. For different values of the characterizing scalar, there now exists a new estimation tool; and
2. The proposed class(es) provides a wide range for selecting the constant scalar by solving the associated polynomials and for root values estimators attains minimum mean squared error.

The proposed methodology is more effective, practicable and efficient, and may be recommended for use in practice.

References

- Cheng, C., & Van Ness, J. W. (1999). *Statistical regression with measurement error*. London, UK: Arnold Publishers.
- Cochran, W. G. (2005). *Sampling techniques*. 2nd Ed. New Delhi, India: Wiley Eastern Private Limited.
- Dalabehara, M., & Sahoo, L. N. (2000). An unbiased estimator in two-phase sampling using two auxiliary variables. *Journal of Indian Society of Agricultural Statistics*, 53(2), 134-140.
- Fuller, W. A. (1987). *Measurement error models*. New York, NY: Wiley.
- Kadilar, C., & Cingi, H. (2005). A new estimator using two auxiliary variables. *Applied Mathematics and Computation*, 162, 901-908.
- Manisha, & Singh, R. K. (2001). Estimation of population mean in the presence of measurement errors. *Journal of Indian Society of Agricultural Statistics*, 54(1), 13-18.
- Mukhopadhyay, P. (2000). *Theory and methods of survey sampling*. New Delhi, India: Prentice Hall of India Pvt. Ltd.
- Murthy, M. N. (1976). *Sampling theory and methods*. Calcutta, IN: Statistical Publishing Society.
- Neter, J. (1970). Measurement errors in reports of consumer expenditures. *Journal of Marketing Research*, 7, 11-25.

Shalabh. (1997). Ratio method of estimation in presence of measurement errors. *Journal of Indian Society of Agricultural Statistics*, 50(2), 150-155.

Shukla, D. (2002). F-T estimator under two-phase sampling. *Metron*, 59(1-2), 253-263.

Shukla, D., Thakur, N. S., Pathak, S., & Rajput, D. S. (2009). Estimation of mean under imputation of missing data using factor-type estimator in two-phase sampling. *Statistics in Transition*, 10(3), 397-414.

Shukla, D., Thakur, N. S. Thakur, D. S., & Pathak, S. (2011). Linear Combination Based Imputation Method for Missing Data in Sample. *International Journal of Modern Engineering Research*, 1(2), 580-596.

Shukla, D., Pathak, S., & Thakur, N. S. (2012). Estimation of population mean using two auxiliary sources in sample surveys. *Statistics in Transition*, 13(1), 21-36.

Shukla, D., Pathak, S., & Thakur, N. S. (2012). An estimator for mean estimation in presence measurement error. *Research & Reviews: A Journal of Statistics*, 1(1), 1-8.

Shukla, D., Pathak, S., & Thakur, N. S. (2012). A transformed estimator for estimation of population mean with missing data in sample-surveys. *Journal of Current Engineering Research*, 2(1), 50-55.

Singh, H. P., & Karpe, N. (2008). Ratio-Product estimator for population mean in presence of measurement errors. *Journal of Applied Statistical Sciences*, 16(4), 49-64.

Singh, V. K., & Shukla, D. (1987). One parameter family of factor-type ratio estimator. *Metron*, 45(1-2), 273-283.

Sud, U. C., Shrivastava, A. K., & Sharma, D. P. (2001). On a biased estimator in repeat surveys. *Journal of the Indian Society of Agricultural Statistics*, 54(1), 29-42.

Sud, V. C., & Shrivastava, A. K. (2000). Estimation and population mean in repeat surveys on the presence of measurement errors. *Journal of Indian Society of Agricultural Statistics*, 53(2), 125-133.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., & Ashok, C. (1984). *Sampling theory of surveys with applications*. New Delhi: Iowa State University Press.

Thakur, N. S., Yadav, K., & Pathak, S. (2011). Estimation of mean in presence of missing data under two-phase sampling scheme. *Journal of Reliability and Statistical Studies*, 4(2), 93-104.

Thakur, N. S., Yadav, K., & Pathak, S. (2012). Some imputation methods in double sampling scheme for estimation of population mean. *International Journal of Modern Engineering Research*, 2(1), 200-207.

Examining Multiple Comparison Procedures According to Error Rate, Power Type and False Discovery Rate

Guven Ozkaya Ilker Ercan
Uludag University,
Gorukle/Bursa, Turkey

Examining pairwise differences between means is a common practice of applied researchers, and the selection of an appropriate multiple comparison procedure (MCP) is important for analyzing pairwise comparisons. This study examines the performance of MCPs under the assumption of homogeneity of variances for various numbers of groups with equal and unequal sample sizes via a simulation study. MCPs are compared according to type I error rate, power type and false discovery rate (FDR). Results show that the LSD and Duncan procedures have high error rates and Scheffe's procedure has low power; no remarkable differences between the other procedures considered were identified.

Key words: Multiple comparison procedures, pairwise comparison, error rates, power, false discovery rate.

Introduction

Multiple comparison procedures (MCPs) are used to test differences between the means of three or more groups after performing variance analysis. Although MCPs are used often, many are not used correctly (Lowry, 1992; Hsu, 1996). Homogeneity of variances, normality and independence of data are assumptions made for variance analysis; these assumptions should also hold when performing MCPs. In addition, sample size also affects MCP performance and should be considered. Some MCPs are purported to apply when the assumptions hold, and some are proposed for the cases in which some assumptions are violated (Demirhan, 2010). Selecting an appropriate MCP is important, it is

necessary to choose a method that is best given the research situation and data. This study examines the performance of MCPs under the assumption of homogeneity of variances for various numbers of groups with equal and unequal sample sizes via a simulation study. MCPs are compared according to type I error rate, power type and false discovery rate (FDR).

General Information

Many MCPs rely on contrasts; a comparison of k groups comprises a comparison of two groups and a comparison of a single group with the remaining groups. This definition is symbolized as

$$\psi = c_1\mu_1 + c_2\mu_2 + \dots + c_k\mu_k = \sum_{j=1}^k c_j\mu_j$$

Guven Ozkaya is a Research Assistant in the Department of Biostatistics, Faculty of Medicine. His research interests are statistical shape analysis and experimental design. Email him at: guvenozkaya@gmail.com. Ilker Ercan is a Professor in the Department of Biostatistics, Faculty of Medicine. His research interests are statistical shape analysis, reliability and validity analyses. Email him at: ercan@uludag.edu.tr.

where the contrast constants $c_1, c_2, \dots, c_k$ sum to zero. MCPs can be categorized according to contrast type as: pairwise comparison, complex comparison, comparison with control and comparison with the best. Generally, the method of contrasts is useful for preplanned, or *a priori*, comparisons, that is, the contrasts are specified prior to conducting the experiment and examining the data. The rationale behind *a priori* contrasts is that, if comparisons are

selected after examining data, many experimenters would construct tests that correspond to large observed differences in means (Montgomery, 2001).

This study focuses on MCPs for examining all possible pairwise comparisons. The LSD, Bonferroni, Dunn-Sidak, Scheffe, REGW-F and Q, Student Neuman Keuls (SNK), Tukey's a and b, Duncan, Hochberg's GT2 and Gabriel procedures are examined for various numbers of groups, variances and sample sizes according to error rate, power type and false discovery rate (FDR).

Measures Used to Evaluate Procedures

The statistical problem that arises from the use of MCPs is that subsequent hypothesis tests will be performed on the outcome with the same data on which the global test was performed: this can result in an uncontrolled type I error rate (Cabral, 2008). However, determining how to control type I errors is much more difficult when multiple significance tests are computed (Jaccard, 2002a, 2002b). This difficulty arises because the decision regarding control of type I errors when MCPs of significance are computed can affect whether the effects are statistically significant (Keselman, 2004).

Choosing from among the various strategies available to control Type I errors could be based on the multiplicity of testing issue. The multiplicity problem in statistical inference refers to the selection of statistically significant findings from a large set of findings (tests) to either support or refute a research hypothesis. Selecting statistically significant findings from a larger pool of results, which also contains non-significant findings, is problematic because when multiple tests of significance are computed the probability that at least one will be significant by chance alone increases with the number of tests examined (Keselman, 2004).

Testing many variables with univariate analysis is typically the first choice for various hypotheses (Tatlidil, 2002), however, due to error rate inflation, this solution is not convenient. There has been much debate concerning the necessity of statistical adjustment

for multiplicity (Kemp, 1975; Bender, 2001). One argument suggests controlling the probability that at least one type I error will occur in the set of pairwise comparison tests by setting that probability equal to alpha (Kemp, 1975; Ludbrook, 1998; Cabral, 2008). This type of control is referred to in the literature as experimentwise or familywise control (Kemp, 1975; Klockars, 1986; Toothaker, 1993; Ludbrook, 1998; Keselman, 2004; Cabral, 2008). In the opposing argument, this type of adjustment is not necessary: instead, each comparison is dealt with separately (O'Neill, 1971; O'Brien, 1983; Perry, 1986; Rothman, 1990). This type of control has been referred to as the comparisonwise error rate (Kemp, 1975; Klockars, 1986; Toothaker, 1993; Keselman, 2004; Cabral, 2008).

The controversy concerning MCPs is whether to control for comparisonwise or experimentwise type I error rates. Related to this controversy is the power of the procedure (Kemp, 1975). It is widely accepted among statisticians that the goal of MCP analysis should be to control the familywise error rate (Keselman, 2004; Toothaker, 1993; Ryan, 1959; Shaffer, 1995; Roback, 2005). Another argument (Benjamini & Hochberg, 1995) is to control the false discovery rate (FDR) (Cabral, 2008; Keselman, 2004; Ludbrook, 1998). Benjamini and Hochberg (1995) developed an alternative approach to multiple hypothesis testing that controls the expected proportion of false positive findings among all rejected hypotheses.

Familywise Error Rate

Familywise error rate is the probability that at least one type I error occurs. Perfect MCPs control the familywise error rate, thus the error rate cannot exceed the α -level (Klockars, 1986; Toothaker, 1993; Ludbrook, 1998; Cabral, 2008) and the type II error rate is minimized. The simultaneous occurrence of two events is impossible (Ludbrook, 1998). MCPs that maintain the overall α -level for a set of tests are said to control the familywise error rate and effectively reduce the α -level for each post hoc test.

EXAMINING MULTIPLE COMPARISON PROCEDURES

Comparisonwise Error Rate

MCPs that apply a separate α -level for each test are called comparisonwise error control procedures (Klockars, 1986; Toothaker, 1993; Cabral, 2008). In a study with groups A, B and C, the use of comparisonwise error control after the global null hypothesis has been rejected entails the performance of 6 individual tests (A-B, A-C, B-C, AB-C, AC-B, BC-A) and the application of an α -level of 0.05 for each test.

False Discovery Rate (FDR)

Much of the debate concerning error rates relates to familywise and comparisonwise error rates. One of the newer interesting contributions to the field of multiple hypothesis testing is an alternative conceptualization for defining errors in the multiple testing problems: the false discovery rate, or FDR, as presented by Benjamini and Hochberg (1995). The FDR is defined by these authors as the expected proportion of the number of erroneous rejections to the total number of rejections. Benjamini and Hochberg (1995) provided several scenarios in which the FDR control seems more reasonable than the familywise or comparisonwise control.

Consider J means, $\mu_1, \mu_2, \dots, \mu_J$, where interest is in testing a family of $m = J(J-1)/2$ pairwise hypotheses, $H_j: \mu_j - \mu_{j'} = 0$ ($j = 1, \dots, J; j' = 1, \dots, J; j \neq j'$), of which m_0 are true. Let S equal the number of correctly rejected hypothesis pairs from the set of R rejections and let the number of falsely rejected pairs be V . Benjamini and Hochberg (1995) summarized the relationship between these random variables (see Table 1). In terms of random variable V , the comparisonwise error rate is $E(V/m)$, whereas the familywise rate is given by $P(V \geq 1)$. Thus, testing each comparison at α guarantees that $E(V/m) \leq \alpha$, whereas testing each comparison at α/m guarantees $P(V \geq 1) \leq \alpha$ (Keselman, 1999). According to Benjamini and Hochberg (1995), the proportion of errors committed by falsely rejecting null hypotheses can be expressed through the random variable $Q = V/(V+S)$, that is, the proportion of rejected hypotheses that are erroneously rejected. It is important to note that Q is defined to be zero when $R = 0$; that is, the error rate is zero when there are no rejections. The FDR was defined by Benjamini and Hochberg as the mean value of Q , that is,

$$\begin{aligned} E(Q) &= E\left(\frac{V}{V+S}\right) \\ &= E\left(\frac{V}{R}\right) \\ &= E\left(\frac{\text{number of false rejections}}{\text{number of total rejections}}\right). \end{aligned}$$

FDR is thus the mean value of the proportion of falsely rejected pairwise tests to the total number of pairwise tests declared significant. As Benjamini and Hochberg indicate, this error rate has a number of important properties:

- a) If $\mu_1 = \mu_2 = \dots = \mu_J$, then all m pairwise comparisons truly equal zero and, therefore, the FDR is equivalent to the familywise error rate; that is, in the case that $s = 0$ and $v = r$, if $v = 0$, then $Q = 0$, and if $V > 0$, then $Q = 1$, and thus $P(V \geq 1) = E(Q)$. Therefore, control of the FDR implies control of the familywise error (Benjamini, 1995).
- b) When $m_0 < m$, the FDR is smaller than or equal to the familywise error rate; in this case, if $v > 1$, then $v/r \leq 1$, and if $V = 0$, then $v/r = 0$ and, thus, $P(V \geq 1) \geq E(Q)$. This result indicates that if the familywise error rate is controlled for a given procedure, then the FDR is also controlled (Keselman, 1999).
- c) v/r tends to be smaller when there are fewer pairs of equal means and when the unequal pairs are more divergent, resulting in a greater difference between the FDR and the familywise value and thus a greater likelihood of increased power by adopting FDR control.

Table 1: Number of Errors Committed when Testing m Null Hypotheses (Benjamini & Hochberg, 1995)

	Declared Non-Significant	Declared Significant	Total
True H_0	U	V	m_0
False H_0	T	S	$m - m_0$
	$m - R$	R	m

Statistical Power Types

The power of MCPs can be categorized for different situations. Over the years, many different conceptualizations of power for (pairwise) comparisons have appeared in the literature. For example, Einot and Gabriel (1975) considered each single subset hypothesis and summarized their findings for all subsets of a particular number of means. Einot and Gabriel (1975) provided results on pair power, triplet power, quadruplet power and quintuplet power. One of the methods to measure power is any-pair power, which is defined as the probability of at least one rejection of a false null hypothesis on a pair of means.

A second measure of power is all-pair power, which is defined as the probability of rejecting all false null hypotheses on pairs of means. If a single false hypothesis is considered, then the probability of rejecting it is called per-pair power. Another power definition considered by Ramsey (Ramsey, 1978; Horn, 2000; Ramsey, 2002; Ramsey & Ramsey, 2008) states that power types are per-pair power, any-pair power, and all-pair power. Ramsey's power types are used frequently in the literature. It should be noted that different names for these terms exist in the literature (Keselman, 2004; Ekenstierna, 2004).

Methodology

The LSD, Bonferroni, Dunn-Sidak, Scheffe, REGW-F and Q, SNK, Tukey's a and b, Duncan, Hochberg's GT2 and Gabriel procedures were examined via simulation scenarios according to error rates, power type and FDR for different numbers of groups, variances and sample sizes. The simulations used 4 cases and 27 scenarios for each case, and 250 replications were made for the 108 scenarios. Data were generated from a normal distribution using R software V.2.11.1 and analyses were performed using SPSS 17.0 for Windows. The four cases examined are:

Case I

Error rates were calculated for equal sample sizes, different numbers of groups and different variances: The data were generated from normal distributions with a mean of 40 and variances 2, 4 and 8. The numbers of compared

groups (k) were 3, 5 and 7. Sample sizes were 10, 30 and 100.

Case II

Error rates were calculated for unequal sample sizes, different numbers of groups and different variances: The data were generated from normal distributions with a mean of 40 and variances 2, 4 and 8. The numbers of compared groups (k) were 3, 5 and 7. Sample sizes were chosen from 10/12/14/16/18/20/22, 30/35/40/45/50/55/60 and 100/110/120/130/140/150/160 for group numbers 3, 5 and 7, respectively.

Case III

Power-type and FDR calculation for equal sample sizes, different numbers of groups and different variances: The data were generated from normal distributions with means of 40/40/42/44/46/48/50 (the first two means are the same due to the FDR calculation) for groups and variances 2, 4 and 8. The numbers of compared groups (k) were 3, 5 and 7. Sample sizes were 10, 30 and 100

Case IV

Power-type and FDR calculation for non-equal sample sizes, different numbers of groups and different variances: The data were generated from normal distributions with means of 40/40/42/44/46/48/50 (the first two means are the same because of the FDR calculation) for groups and variances 2, 4 and 8. The numbers of compared groups (k) were 3, 5 and 7. Sample sizes were chosen from 10/12/14/16/18/20/22, 30/35/40/45/50/55/60 and 100/110/120/130/140/150/160 for group sizes 3, 5, and 7, respectively.

Results

Simulation results for error rates are shown in Tables 2-5, and the power-type and FDR results are summarized in Table 6. When the number of groups is small, for both equal and unequal sample sizes, the LSD and Duncan error rates are higher than the other MCPs; the other MCP error rates are very similar. Although the number of groups is increasing, the familywise error rates are highest with the LSD and Duncan procedures and are lowest with the Scheffe. In addition, the comparisonwise error rates of the LSD and Duncan procedures are the highest and

EXAMINING MULTIPLE COMPARISON PROCEDURES

the Bonferroni, Dunn-Sidak and Scheffe error rates are the lowest for both equal and unequal sample sizes.

For a small number of groups and equal and unequal sample sizes, the LSD, Duncan and REGW-F procedures have the highest per-pair power and all-pair power. As the number of groups increases, the LSD, Duncan and SNK procedures have the highest per-pair power. The Scheffe per-pair power is the lowest among all the MCPs. For a large sample size, the per-pair power and all-pair power of all the MCPs are very close.

The three highest MCP any-pair powers are those for the LSD, Duncan and REGW-F procedures for small groups. As the number of groups increases, all the MCP powers reach their highest values for equal and unequal sample sizes.

For small groups and equal and unequal sample sizes, the LSD, REGW-F, REGW-Q, SNK, Tukey's b and Duncan procedures have FDR values higher than those of the other procedures. As the number of groups increases, the FDR values of all MCPs become similar. Also, as the number of comparisons increases, the FDR decreases.

Discussion

MCPs were studied in terms of the familywise error rate for equal and unequal sample sizes; the Duncan's and the LSD's familywise error rates were very high. For both procedures, when the number of groups increased, the familywise error rate also increased. For a large number of groups, the Scheffe procedure has the lowest familywise error rate of all the MCPs and was not affected by the change in the-sample sizes.

The familywise error rates of the Bonferroni, Dunn-Sidak, Gabriel and Hochberg's GT2 procedures were low and similar to each other for small numbers of groups. When the number of groups was increased, the familywise error rates of the LSD and Duncan procedures were the highest. The Scheffe procedure was the lowest of all the MCPs.

The FWE values of REGW-F, REGW-Q, SNK, Tukey's a and b, Gabriel and Hochberg's GT2 procedures were similar.

There was no significant change in the MCP familywise error rates due to an increase or decrease in number of groups, with the exceptions of the LSD and Duncan procedures. There also was no significant change in the familywise error rate for any MCP due to changes in sample size. Based on the homogeneity of variance assumption, changes in variance were not considered to have a serious an effect on familywise error rate.

According to comparisonwise error rate results for equal and unequal sample sizes, the familywise error rates of the LSD and the Duncan procedures were higher than those of other procedures. The comparisonwise error rate of the LSD procedure was not greatly affected by changes in group number. Conversely, the Duncan comparisonwise error rate significantly increased with increases in group number. For equal and unequal sample sizes, the Bonferroni, Dunn-Sidak and Scheffe procedures generally had the lowest comparisonwise error rates.

There were no significant changes in the comparisonwise error rates of these procedures due to increases in group number. The comparisonwise error rates of the REGW-F, REGW-Q, SNK, Tukey's a and b, Gabriel and Hochberg's GT2 procedures were not as low as those of the Bonferroni, Dunn-Sidak and Scheffe procedures, however, they were lower than those of the LSD and Duncan procedures. The comparisonwise REGW-Q error rate was lower than the LSD error rate. There were no significant changes in comparisonwise error rate because of the increases in group number; in addition, there was no significant change in the comparisonwise error rate of any MCP because of the changes in sample size and variance.

The per-pair power of all MCPs increased as variance decreased and as sample size increased for equal and unequal sample sizes. Furthermore, the per-pair powers of all MCPs increased with the increase of group number. The LSD, Duncan and REGW-F procedures had the highest per-pair power for a small number of groups and a small sample size.

With larger group numbers and sample sizes, the LSD, Duncan and SNK procedures had the highest per-pair powers and the Scheffe

Table 2: Familywise Error Rates Results for Equal Sample Size

	Number of Groups 3								
	$n_i: 10/10/10$			$n_i: 30/30/30$			$n_i: 100/100/100$		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.116	0.108	0.120	0.136	0.096	0.088	0.080	0.132	0.128
Bonferroni	0.020	0.028	0.032	0.036	0.036	0.036	0.028	0.040	0.036
Dunn-Sidak	0.020	0.028	0.032	0.036	0.036	0.036	0.028	0.040	0.036
Scheffe	0.020	0.028	0.028	0.036	0.032	0.032	0.020	0.036	0.028
REGW - F	0.032	0.032	0.040	0.036	0.040	0.040	0.032	0.044	0.032
REGW - Q	0.024	0.032	0.048	0.036	0.036	0.040	0.040	0.044	0.048
SNK	0.028	0.032	0.048	0.036	0.036	0.040	0.040	0.044	0.048
Tukey a	0.024	0.036	0.052	0.040	0.040	0.044	0.044	0.048	0.056
Tukey b	0.028	0.032	0.048	0.036	0.036	0.040	0.040	0.044	0.048
Duncan	0.076	0.072	0.096	0.092	0.068	0.076	0.056	0.104	0.092
Hochberg's GT2	0.020	0.028	0.040	0.036	0.036	0.036	0.028	0.044	0.036
Gabriel	0.020	0.028	0.040	0.036	0.036	0.036	0.028	0.044	0.036

	Number of Groups 5								
	$n_i: 10/10/10/10/10$			$n_i: 30/30/30/30/30$			$n_i: 100/100/100/100/100$		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.240	0.196	0.196	0.212	0.216	0.188	0.244	0.184	0.236
Bonferroni	0.032	0.028	0.028	0.028	0.044	0.032	0.032	0.024	0.044
Dunn-Sidak	0.032	0.028	0.028	0.028	0.044	0.032	0.032	0.032	0.048
Scheffe	0.024	0.016	0.012	0.004	0.020	0.008	0.008	0.012	0.016
REGW - F	0.040	0.040	0.032	0.060	0.064	0.048	0.048	0.032	0.052
REGW - Q	0.040	0.044	0.032	0.032	0.052	0.052	0.044	0.032	0.048
SNK	0.040	0.044	0.032	0.032	0.052	0.056	0.044	0.032	0.048
Tukey a	0.040	0.044	0.032	0.032	0.052	0.052	0.044	0.032	0.048
Tukey b	0.040	0.044	0.032	0.032	0.052	0.056	0.044	0.032	0.048
Duncan	0.168	0.132	0.136	0.160	0.148	0.140	0.168	0.132	0.160
Hochberg's GT2	0.044	0.040	0.036	0.032	0.044	0.032	0.040	0.036	0.048
Gabriel	0.032	0.032	0.028	0.028	0.044	0.032	0.032	0.032	0.048

	Number of Groups 7								
	$n_i: 10/10/10/10/10/10/10$			$n_i: 30/30/30/30/30/30/30$			$n_i: 100/100/100/100/100/100/100$		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.436	0.436	0.464	0.424	0.420	0.408	0.436	0.432	0.400
Bonferroni	0.036	0.024	0.040	0.044	0.032	0.064	0.044	0.024	0.052
Dunn-Sidak	0.036	0.028	0.040	0.044	0.032	0.068	0.044	0.028	0.052
Scheffe	0.008	0.004	0.016	0.016	0.012	0.016	0.004	0.004	0.008
REGW - F	0.048	0.028	0.044	0.048	0.032	0.048	0.052	0.036	0.044
REGW - Q	0.048	0.020	0.036	0.048	0.040	0.068	0.044	0.028	0.048
SNK	0.048	0.020	0.036	0.048	0.040	0.068	0.044	0.028	0.048
Tukey a	0.052	0.040	0.056	0.052	0.052	0.076	0.052	0.040	0.060
Tukey b	0.048	0.020	0.036	0.048	0.044	0.068	0.044	0.028	0.048
Duncan	0.220	0.184	0.248	0.232	0.216	0.208	0.248	0.200	0.196
Hochberg's GT2	0.036	0.028	0.040	0.044	0.032	0.072	0.044	0.028	0.052
Gabriel	0.036	0.028	0.040	0.044	0.036	0.072	0.044	0.028	0.052

EXAMINING MULTIPLE COMPARISON PROCEDURES

Table 3: Familywise Error Rates Results for Unequal Sample Size

	Number of Groups 3								
	n_i :10/12/14			n_i :30/35/40			n_i :100/110/120		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.132	0.124	0.084	0.124	0.172	0.140	0.116	0.080	0.128
Bonferroni	0.024	0.036	0.020	0.044	0.056	0.048	0.052	0.024	0.024
Dunn-Sidak	0.028	0.036	0.020	0.044	0.056	0.048	0.052	0.024	0.024
Scheffe	0.020	0.032	0.008	0.036	0.056	0.040	0.044	0.020	0.024
REGW - F	0.028	0.048	0.024	0.056	0.068	0.044	0.044	0.036	0.028
REGW - Q	0.024	0.032	0.016	0.036	0.052	0.044	0.044	0.020	0.024
SNK	0.036	0.032	0.020	0.056	0.068	0.052	0.052	0.032	0.024
Tukey a	0.044	0.040	0.028	0.060	0.076	0.048	0.060	0.036	0.028
Tukey b	0.040	0.032	0.020	0.056	0.068	0.052	0.052	0.032	0.024
Duncan	0.092	0.088	0.056	0.092	0.128	0.108	0.080	0.060	0.084
Hochberg's GT2	0.032	0.032	0.020	0.044	0.064	0.044	0.048	0.024	0.024
Gabriel	0.032	0.032	0.020	0.044	0.064	0.044	0.048	0.024	0.024

	Number of Groups 5								
	n_i :10/12/14/16/18			n_i :30/35/40/45/50			n_i :100/110/120/130/140		
LSD	0.240	0.188	0.200	0.188	0.240	0.280	0.224	0.224	0.240
Bonferroni	0.032	0.040	0.024	0.028	0.032	0.036	0.024	0.044	0.032
Dunn-Sidak	0.032	0.040	0.024	0.028	0.032	0.040	0.024	0.044	0.032
Scheffe	0.012	0.024	0.012	0.016	0.000	0.016	0.012	0.024	0.012
REGW - F	0.032	0.036	0.028	0.032	0.036	0.040	0.028	0.044	0.040
REGW - Q	0.020	0.032	0.016	0.024	0.020	0.020	0.020	0.044	0.024
SNK	0.028	0.032	0.020	0.032	0.036	0.056	0.032	0.056	0.032
Tukey a	0.036	0.036	0.024	0.036	0.028	0.052	0.032	0.056	0.032
Tukey b	0.028	0.032	0.020	0.032	0.036	0.056	0.032	0.056	0.032
Duncan	0.136	0.128	0.128	0.128	0.152	0.204	0.140	0.164	0.152
Hochberg's GT2	0.036	0.028	0.024	0.028	0.032	0.044	0.028	0.044	0.036
Gabriel	0.032	0.028	0.024	0.028	0.028	0.040	0.028	0.044	0.032

	Number of Groups 7								
	n_i :10/12/14/16/18/20/22			n_i :30/35/40/45/50/55/60			n_i :100/110/120/130/140/150/160		
LSD	0.396	0.444	0.488	0.436	0.408	0.432	0.444	0.408	0.468
Bonferroni	0.024	0.044	0.040	0.040	0.012	0.032	0.024	0.028	0.036
Dunn-Sidak	0.024	0.044	0.044	0.040	0.016	0.032	0.024	0.028	0.036
Scheffe	0.012	0.020	0.012	0.012	0.004	0.000	0.000	0.012	0.004
REGW - F	0.048	0.048	0.044	0.032	0.028	0.044	0.032	0.032	0.036
REGW - Q	0.020	0.032	0.012	0.028	0.020	0.024	0.012	0.020	0.024
SNK	0.040	0.052	0.068	0.052	0.020	0.036	0.028	0.028	0.044
Tukey a	0.036	0.056	0.100	0.068	0.024	0.036	0.044	0.036	0.052
Tukey b	0.040	0.052	0.068	0.052	0.020	0.036	0.028	0.028	0.044
Duncan	0.204	0.244	0.232	0.228	0.216	0.224	0.196	0.204	0.212
Hochberg's GT2	0.032	0.044	0.084	0.044	0.016	0.032	0.024	0.028	0.040
Gabriel	0.032	0.044	0.080	0.044	0.016	0.032	0.024	0.028	0.040

Table 4: Comparisonwise Error Rates Results for Equal Sample Size

	Number of Groups 3								
	$n_i: 10/10/10$			$n_i: 30/30/30$			$n_i: 100/100/100$		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.048	0.048	0.055	0.057	0.048	0.039	0.040	0.063	0.061
Bonferroni	0.011	0.011	0.011	0.016	0.021	0.016	0.015	0.015	0.017
Dunn-Sidak	0.011	0.011	0.011	0.016	0.021	0.016	0.015	0.015	0.017
Scheffe	0.011	0.011	0.009	0.015	0.019	0.015	0.009	0.012	0.013
REGW - F	0.013	0.016	0.020	0.020	0.024	0.019	0.016	0.024	0.019
REGW - Q	0.009	0.016	0.023	0.019	0.020	0.017	0.019	0.023	0.024
SNK	0.012	0.016	0.023	0.019	0.020	0.017	0.019	0.023	0.024
Tukey a	0.012	0.015	0.019	0.020	0.023	0.019	0.020	0.017	0.025
Tukey b	0.011	0.016	0.020	0.017	0.019	0.017	0.017	0.019	0.021
Duncan	0.029	0.029	0.039	0.040	0.033	0.031	0.024	0.045	0.039
Hochberg's GT2	0.011	0.011	0.013	0.016	0.021	0.016	0.015	0.015	0.017
Gabriel	0.011	0.011	0.013	0.016	0.021	0.016	0.015	0.015	0.017

	Number of Groups 5								
	$n_i: 10/10/10/10/10$			$n_i: 30/30/30/30/30$			$n_i: 100/100/100/100/100$		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.058	0.050	0.048	0.053	0.048	0.044	0.048	0.042	0.055
Bonferroni	0.004	0.006	0.007	0.005	0.007	0.006	0.004	0.004	0.008
Dunn-Sidak	0.004	0.006	0.007	0.005	0.007	0.006	0.004	0.004	0.008
Scheffe	0.005	0.004	0.006	0.002	0.008	0.002	0.002	0.002	0.003
REGW - F	0.013	0.010	0.008	0.022	0.031	0.018	0.016	0.007	0.014
REGW - Q	0.009	0.012	0.008	0.016	0.027	0.020	0.013	0.007	0.012
SNK	0.012	0.012	0.008	0.016	0.027	0.022	0.014	0.007	0.013
Tukey a	0.010	0.013	0.012	0.017	0.021	0.020	0.012	0.008	0.013
Tukey b	0.009	0.011	0.008	0.015	0.027	0.020	0.013	0.007	0.012
Duncan	0.060	0.054	0.042	0.070	0.069	0.057	0.065	0.044	0.059
Hochberg's GT2	0.008	0.011	0.012	0.016	0.018	0.012	0.011	0.009	0.013
Gabriel	0.006	0.010	0.010	0.014	0.018	0.012	0.009	0.007	0.012

	Number of Groups 7								
	$n_i: 10/10/10/10/10/10/10$			$n_i: 30/30/30/30/30/30/30$			$n_i: 100/100/100/100/100/100/100$		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.052	0.047	0.057	0.054	0.045	0.050	0.053	0.049	0.050
Bonferroni	0.002	0.001	0.003	0.003	0.002	0.004	0.003	0.001	0.004
Dunn-Sidak	0.002	0.001	0.003	0.003	0.002	0.004	0.003	0.002	0.005
Scheffe	0.003	0.003	0.002	0.003	0.004	0.004	0.000	0.000	0.005
REGW - F	0.014	0.010	0.014	0.018	0.012	0.020	0.022	0.012	0.023
REGW - Q	0.013	0.007	0.010	0.013	0.014	0.022	0.018	0.005	0.024
SNK	0.016	0.010	0.010	0.013	0.016	0.026	0.018	0.005	0.028
Tukey a	0.013	0.008	0.011	0.013	0.030	0.018	0.012	0.006	0.022
Tukey b	0.013	0.007	0.009	0.013	0.018	0.022	0.017	0.005	0.024
Duncan	0.106	0.074	0.100	0.114	0.095	0.083	0.097	0.083	0.099
Hochberg's GT2	0.012	0.007	0.007	0.011	0.011	0.018	0.012	0.004	0.016
Gabriel	0.012	0.007	0.007	0.011	0.016	0.018	0.012	0.004	0.016

EXAMINING MULTIPLE COMPARISON PROCEDURES

Table 5: Comparisonwise Error Rates Results for Unequal Sample Size

	Number of Groups 3								
	n_i : 10/12/14			n_i : 30/35/40			n_i : 100/110/120		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.051	0.053	0.033	0.049	0.075	0.075	0.052	0.045	0.052
Bonferroni	0.011	0.017	0.007	0.015	0.024	0.024	0.019	0.011	0.008
Dunn-Sidak	0.012	0.017	0.007	0.016	0.024	0.025	0.019	0.011	0.008
Scheffe	0.009	0.016	0.003	0.012	$\bar{\Gamma} \vee \hat{H}$	0.021	0.016	0.009	0.008
REGW - F	0.013	0.023	0.012	0.025	0.033	0.024	0.024	0.019	0.012
REGW - Q	0.011	0.015	0.008	0.013	0.020	0.024	0.024	0.009	0.009
SNK	0.015	0.015	0.009	0.025	0.028	0.028	0.027	0.016	0.011
Tukey a	0.017	0.019	0.011	0.021	0.031	0.027	0.021	0.015	0.009
Tukey b	0.016	0.013	0.008	0.021	0.024	0.024	0.023	0.011	0.009
Duncan	0.035	0.037	0.023	0.039	0.052	0.051	0.037	0.027	0.032
Hochberg's GT2	0.013	0.016	0.007	0.016	0.027	0.024	0.017	0.011	0.008
Gabriel	0.013	0.016	0.007	0.016	0.027	0.024	0.017	0.011	0.008

	Number of Groups 5								
	n_i : 10/12/14/16/18			n_i : 30/35/40/45/50			n_i : 100/110/120/130/140		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.053	0.047	0.043	0.047	0.054	0.057	0.050	0.050	0.054
Bonferroni	0.006	0.006	0.004	0.005	0.006	0.006	0.003	0.007	0.006
Dunn-Sidak	0.006	0.006	0.004	0.005	0.006	0.006	0.003	0.007	0.006
Scheffe	0.004	0.005	0.005	0.007	0.001	0.006	0.004	0.006	0.003
REGW - F	0.013	0.016	0.010	0.011	0.009	0.016	0.010	0.013	0.012
REGW - Q	0.006	0.010	0.006	0.010	0.005	0.008	0.008	0.009	0.010
SNK	0.012	0.016	0.012	0.014	0.010	0.017	0.012	0.012	0.011
Tukey a	0.014	0.010	0.010	0.014	0.008	0.015	0.010	0.013	0.009
Tukey b	0.011	0.012	0.010	0.012	0.010	0.017	0.012	0.012	0.011
Duncan	0.054	0.057	0.043	0.059	0.053	0.065	0.062	0.055	0.048
Hochberg's GT2	0.014	0.008	0.010	0.013	0.009	0.014	0.010	0.012	0.011
Gabriel	0.013	0.007	0.009	0.012	0.007	0.012	0.009	0.010	0.009

	Number of Groups 7								
	n_i : 10/12/14/16/18/20/22			n_i : 30/35/40/45/50/55/60			n_i : 100/110/120/130/140/150/160		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.050	0.052	0.063	0.049	0.040	0.046	0.049	0.046	0.055
Bonferroni	0.002	0.002	0.003	0.003	0.001	0.002	0.002	0.002	0.002
Dunn-Sidak	0.002	0.002	0.003	0.004	0.001	0.002	0.002	0.002	0.002
Scheffe	0.005	0.003	0.002	0.002	0.000	0.000	0.000	0.005	0.000
REGW - F	0.017	0.013	0.007	0.006	0.006	0.013	0.013	0.012	0.012
REGW - Q	0.008	0.008	0.002	0.005	0.005	0.011	0.002	0.010	0.005
SNK	0.024	0.021	0.018	0.011	0.005	0.013	0.005	0.018	0.008
Tukey a	0.016	0.014	0.016	0.011	0.004	0.012	0.006	0.014	0.007
Tukey b	0.021	0.017	0.016	0.010	0.004	0.013	0.005	0.014	0.008
Duncan	0.115	0.097	0.084	0.083	0.071	0.083	0.093	0.078	0.086
Hochberg's GT2	0.012	0.012	0.014	0.008	0.003	0.010	0.002	0.010	0.005
Gabriel	0.012	0.012	0.014	0.008	0.003	0.010	0.002	0.010	0.005

Table 6: Power and FDR Results of MCPs According to Variance, Number of Groups and Sample Size*

n	σ^2	Per-Pair Power			Any-Pair Power		
		k=3	k=5	k=7	k=3	k=5	k=7
$n_i = \dots = n_j = 10$	2	.712-.854	.769-.942	.795-.961	.864-.932	1.00-1.00	1.00-1.00
	4	.366-.566	.598-.817	.674-.879	.492-.696	1.00-1.00	1.00-1.00
	8	.180-.330	.397-.666	.504-.771	.284-.476	1.00-1.00	1.00-1.00
$n_i = \dots = n_j = 30$	2	.996-.996	.996-1.00	.991-1.00	.996-.996	1.00-1.00	1.00-1.00
	4	.898-.960	.889-.987	.881-.993	.968-.996	1.00-1.00	1.00-1.00
	8	.614-.774	.702-.894	.753-.932	.752-.876	1.00-1.00	1.00-1.00
$n_i = \dots = n_j = 100$	2	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	4	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	8	.992-1.00	.987-.999	.978-1.00	1.00-1.00	1.00-1.00	1.00-1.00
$n_i \neq \dots \neq n_j$ 10/12/14/16/ 18/20/22	2	.858-.932	.873-.977	.892-.993	.948-.980	1.00-1.00	1.00-1.00
	4	.444-.664	.669-.886	.762-.939	.620-.820	1.00-1.00	1.00-1.00
	8	.246-.432	.519-.746	.624-.841	.380-.588	.996-1.00	1.00-1.00
$n_i \neq \dots \neq n_j$ 30/35/40/45/ 50/55/60	2	1.00-1.00	.999-1.00	.999-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	4	.968-.990	.946-.996	.953-.999	1.00-1.00	1.00-1.00	1.00-1.00
	8	.732-.866	.789-.949	.821-.970	.864-.948	1.00-1.00	1.00-1.00
$n_i \neq \dots \neq n_j$ 100/110/120/ 130/140/150/160	2	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	4	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	8	.996-1.00	.996-1.00	.993-1.00	1.00-1.00	1.00-1.00	1.00-1.00

*Results are summarized with minimum and maximum values (min-max)

procedure has the lowest per-pair power. The per-pair power of the REGW-Q procedure was not as powerful as the LSD or Duncan procedures; also, the SNK per-pair power was similar to the LSD and Duncan procedures. The power of the Bonferroni, Dunn-Sidak, Gabriel and Hochberg GT2 procedures were very close and the REGW-F, REGW-Q, and Tukey's a and b procedures are close for equal and unequal sample sizes.

Like per-pair, any-pair and all-pair powers increased as variance decreased and as sample size increased. If the group number and sample sizes were small, the LSD, Duncan and REGW-F procedures had high any-pair powers and the Scheffe procedure had the lowest any-pair power. As the number of groups increased, any-pair power reached its highest level.

All-pair-power decreased as the number of groups and the variance both increased. When sample size and number of groups were small,

EXAMINING MULTIPLE COMPARISON PROCEDURES

Table 6 (continued): Power and FDR Results of MCPs According to Variance, Number of Groups and Sample Size*

n	σ^2	All-Pair-Power			False Discovery Rate		
		k=3	k=5	k=3	k=5	k=3	k=5
$n_i = \dots = n_j = 10$	2	.560-.776	.044-.564	.560-.776	.044-.564	.560-.776	.044-.564
	4	.240-.436	.000-.092	.240-.436	.000-.092	.240-.436	.000-.092
	8	.076-.196	.000-.004	.076-.196	.000-.004	.076-.196	.000-.004
$n_i = \dots = n_j = 30$	2	.996-.996	.968-1.00	.996-.996	.968-1.00	.996-.996	.968-1.00
	4	.828-.924	.276-.884	.828-.924	.276-.884	.828-.924	.276-.884
	8	.476-.672	.000-.292	.476-.672	.000-.292	.476-.672	.000-.292
$n_i = \dots = n_j = 100$	2	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	4	1.00-1.00	.996-1.00	1.00-1.00	.996-1.00	1.00-1.00	.996-1.00
	8	.984-1.00	.888-.992	.984-1.00	.888-.992	.984-1.00	.888-.992
$n_i \neq \dots \neq n_j$ 10/12/14/16/ 18/20/22	2	.768-.884	.232-.812	.768-.884	.232-.812	.768-.884	.232-.812
	4	.264-.508	.000-.284	.264-.508	.000-.284	.264-.508	.000-.284
	8	.112-.296	.000-.012	.112-.296	.000-.012	.112-.296	.000-.012
$n_i \neq \dots \neq n_j$ 30/35/40/45/ 50/55/60	2	1.00-1.00	.992-1.00	1.00-1.00	.992-1.00	1.00-1.00	.992-1.00
	4	.936-.980	.616-.964	.936-.980	.616-.964	.936-.980	.616-.964
	8	.592-.784	.044-.592	.592-.784	.044-.592	.592-.784	.044-.592
$n_i \neq \dots \neq n_j$ 100/110/120/ 130/140/150/160	2	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	4	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	8	.992-1.00	.968-.100	.992-1.00	.968-.100	.992-1.00	.968-.100

*Results are summarized with minimum and maximum values (min-max)

the LSD, Duncan and REGW-F procedures had the highest power and the Scheffe procedure had the lowest power. As the number of groups increased, the LSD, Duncan and SNK procedures had the highest all-pair powers and the Scheffe procedure has the lowest all-pair power for both equal and unequal sample sizes.

For small groups and equal and unequal sample sizes, the LSD, REGW-F, REGW-Q, SNK, Tukey's b and Duncan FDR values were higher than the other values. As the number of

groups increased all MCPs become closer and, as the number of comparisons increased, the FDR gets smaller.

The familywise error rates of the SNK and REGW-Q procedures were not as high as those of the LSD and Duncan procedures. Similarly, for the comparisonwise error rate, the LSD and Duncan procedures had the highest rates, whereas the Scheffe, Bonferroni and Dunn-Sidak procedures had the smallest rates. The LSD, Duncan and SNK procedures had the

highest powers, whereas the Scheffe procedure had the lowest power.

The any-pair powers of the LSD and Duncan procedures were high, but the Scheffe power was low due to small sample size and number of groups; power reached its maximum value as the number of groups increased. The all-pair powers of the LSD, Duncan, REGW-F, REGW-Q, SNK and Tukey's a and b procedures were the highest, but the Scheffe power was the lowest due to small sample size and number of groups. As the number of groups increased, the LSD, Duncan and SNK procedures had the highest power and the Scheffe procedure had the lowest power. FDR values of the LSD, REGW-F, REGW-Q, SNK and Duncan procedures were higher than those of the other procedures for low number of groups.

Conclusion

Findings from this study that differed from the literature were: (1) the SNK procedure is as robust as the LSD and Duncan procedures for controlling the error rate (Bernhardson, 1975; Curran-Everett, 2000; Maxwell, 2004), and (2) the REGW-Q procedure is as robust as the LSD for CWE (Menéndez De La Fuente, 1999).

Based on study results, the LSD and Duncan procedures are not recommended due to high error rates. The Scheffe procedure is not recommended due to its low power. There were no remarkable differences between the other procedures, thus, it is not possible to recommend one specific pairwise MCP for all situations that applied researchers may encounter.

References

- Bender, R., & Lange, S. (2001). Adjusting for multiple testing-when and how? *Journal of Clinical Epidemiology*, 54(4), 343-349.
- Benjamini, Y., & Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289-300.
- Bernhardson, C. (1975). Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. *Biometrics*, 31, 229-232.
- Cabral, H. J. (2008). Multiple comparisons procedures. *Circulation*, 117(5), 698-701.
- Curran-Everett, D. (2000). Multiple comparisons: philosophies and illustrations. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 279(1), R1-R8.
- Dayton, C. M. (1998). Information criteria for the paired-comparisons problem. *The American Statistician*, 52(2), 144-151.
- Demirhan, H., Dolgun, N. A., Demirhan, Y. P., & Dolgun, M. O. (2010). Performance of some multiple comparison tests under heteroscedasticity and dependency. *Journal of Statistical Computation and Simulation*, 80(10), 1083-1100.
- Einot, I., & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, 70(351), 574-583.
- Ekenstierna, M. (2004). *Multiple comparison procedures based on marginal p-values*. Uppsala University Department of Mathematics Project Report No: 2004(12), 1-26. <http://www2.math.uu.se/research/pub/Ekenstierna.pdf>.
- Horn, M., & Vollandt, R. (2000). A survey of sample size formulas for pairwise and many one multiple comparisons in the parametric, nonparametric and binomial Case. *Biometrical Journal*, 42(1), 27-44.
- Hsu, J. C. (1996). *Multiple comparisons: theory and methods*. London, England: Chapman & Hall/CRC.
- Jaccard, J., & Guilamo-Ramos, V. (2002a). Analysis of variance frameworks in clinical child and adolescent psychology: Issues and recommendations. *Journal of Clinical Child and Adolescent Psychology*, 31, 130-146.
- Jaccard, J., & Guilamo-Ramos V. (2002b). Analysis of variance frameworks in clinical child and adolescent psychology: Advanced issues and recommendations. *Journal of Clinical Child and Adolescent Psychology*, 31, 278-294.
- Kemp, K. (1975). Multiple comparisons: Comparisonwise versus experimentwise type I error rates and their relationship to power. *Journal of Dairy Science*, 58(9), 1374-1378.

EXAMINING MULTIPLE COMPARISON PROCEDURES

- Keselman, H., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparisonwise Type I error control. *Psychological Methods*, 4, 58-69.
- Keselman, H., Cribbie, R. A., & Holland, B. (2004). Pairwise multiple comparison test procedures: an update for clinical child and adolescent psychologists. *Journal of Clinical Child & Adolescent Psychology*, 33(3), 623-645.
- Keselman, H., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3), 350.
- Klockars, A. J., & Sax, G. (1986). *Multiple comparisons*. Newbury Park, CA: Sage Publications.
- Lowry, S. (1992). Use and misuse of multiple comparisons in animal experiments. *Journal of Animal Science*, 70(6), 1971-1977.
- Ludbrook, J. (1998). Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology*, 25(12), 1032-1037.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. London, England: Lawrence Erlbaum Associates.
- Menéndez De La Fuente, I., Vallejo Seco, G., & Fernández García P. (1999). Multiple comparison procedures for simple one-way ANOVA with dependent data. *Spanish journal of psychology*, 2(1), 55-63.
- Montgomery, D. C. (2001). *Design and analysis of experiments*. New York, NY: John Wiley & Sons.
- O'Brien, P. (1983). The appropriateness of analysis of variance and multiple-comparison procedures. *Biometrics*, 39(3), 787-788.
- O'Neill, R., & Wetherill, G. (1971). The present state of multiple comparison methods. *Journal of the Royal Statistical Society, Series B (Methodological)*, 33(2), 218-250.
- Perry, J. (1986). Multiple-comparison procedures: a dissenting view. *Journal of Economic Entomology*, 79(5), 1149-1155.
- Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, 73(363), 479-485.
- Ramsey, P. H. (2002). Comparison of closed testing procedures for pairwise testing of means. *Psychological Methods*, 7(4), 504-523.
- Ramsey, P. H., & Ramsey P. P. (2008). Power of pairwise comparisons in the equal variance and unequal sample size case. *British Journal of Mathematical and Statistical Psychology*, 61(1), 115-131.
- Roback, P. J., & Askins, R. A. (2005). Judicious use of multiple hypothesis tests. *Conservation Biology*, 19(1), 261-267.
- Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, 1(1), 43-46.
- Ryan, T. A. (1959). Multiple comparison in psychological research. *Psychological Bulletin*, 56(1), 26-47.
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561-584.
- Tatlidil, H. (2002). *Applied Multivariate Statistical Analysis*, Ziraat Matbaacılık, Ankara. (in Turkish).
- Toothaker, L. E. (1993). *Multiple comparison procedures*. Newbury Park, CA: Sage Publications.

Selection of Mixed Sampling Plan with QSS-1(n ; c_N , c_T) Plan as Attribute Plan Indexed Through MAPD and LQL

R. Sampath Kumar
Government Arts College,
Coimbatore, India

M. Indra
Muthayammal College of
Arts & Science,
Rasipuram, India

R. Radhakrishnan
PSG College of Arts & Science,
Coimbatore, India

A procedure for the construction and selection of the mixed sampling plan using MAPD as a quality standard with the QSS-1 (n ; c_N , c_T) plan as an attribute plan is presented. The plans indexed through MAPD and LQL are constructed and compared for efficiency. Tables are provided for selection of an appropriate sampling plan.

Key words: Limiting quality level, maximum allowable percent defective, operating characteristic, tangent intercept.

Introduction

Mixed sampling plans consist of two stages with different natures. During the first stage a given lot is considered as a sample from the respective production process and a criterion by variables is used to check process quality. If process quality is judged to be sufficiently good then the lot is accepted, if not, the second stage of the

sampling plan is entered and lot quality is checked directly by means of an attribute sampling plan. There are two types of mixed sampling plans called independent and dependent plans. If the first stage sample results are not utilized in the second stage, then the plan is said to be independent, otherwise it is considered to be dependent. The principal advantage of a mixed sampling plan over a pure attribute sampling plan is a reduction in sample size for a similar amount of protection.

The second stage attribute inspection becomes more important to discriminate the lot if the first stage variable inspection fails to accept the lot. If rejection occurs during the normal inspection, tightened inspection is recommended in the mixed system and vice versa in the second stage. Hence Quick Switching System is imposed in the second stage to sharpen the sampling situation and to insist the producer to manufacture goods within the Limiting Quality Level. Dodge (1967) proposed a sampling system called a 'Quick Switching System' (QSS) consisting of pairs of normal and tightened plans.

Schilling (1967) proposed a method for determining the operating characteristics of mixed variables – attributes sampling plans, single sided specification and standard deviation known using the normal approximation.

R. Sampath Kumar is an Assistant Professor in the Department of Statistics. He has 12 years of teaching experience in theoretical and applied Statistics, has presented 13 papers in both national and international conferences and has published 10 articles in reputed national and international journals. Email him at: rsampath1969@gmail.com. M. Indra is an Assistant Professor in the Department of Statistics. She has been teaching statistics for 7 years. Email: indra_m1980@rediffmail.com. R. Radhakrishnan is an Associate Professor in the Department of Statistics. He has 31 years of experience teaching theoretical and applied statistics, has presented more than 150 papers at national and international conferences and has published more than 100 articles. He is a quality auditor for ISO certification and a certified Six Sigma Black Belt. Email him at: rkrishnan_research@yahoo.com.

Devaarul (2003), Radhakrishnan and Sampath Kumar (2006a, 2006b, 2007a, 2007b, 2009) have investigated mixed sampling plans for the independent case. Radhakrishnan, et al. (2009) studied mixed sampling plan for the dependent case. Quick Switching System (QSS) were originally proposed by Dodge (1967) and have been investigated by Romboski (1969) and Govindaraju (1991). Dodge (1967) proposed a new sampling system consisting of pairs of normal and tightened plans. Romboski (1969) developed a QSS by attributes with a reduction in the sample size required to achieve approximately the same operating characteristic curve.

This study uses the operating procedure of mixed sampling plan with a QSS-1 ($n;c_N,c_T$) plan as an attribute plan to construct tables for a mixed sampling plan indexed through (i) maximum allowable percent defective (MAPD), and (ii) limiting quality level (LQL). The plan indexed through MAPD is compared to the plan indexed through LQL.

Conditions of Applications of QSS-1-Mixed Sampling Plan

The following assumptions are made with respect to the application conditions of a QSS-1 mixed sampling plan:

- Production is steady so that results regarding current and preceding lots are broadly indicative of a continuing process.
- Lots are submitted substantially in the order of their production.
- Inspection involves costly or destructive tests such that normally only a small number of tests per lot can be justified.

Glossary of Symbols

The symbols used in this article are:

p : submitted quality of lot or process;

$P_a(p)$: probability of acceptance for given quality ' p ';

P_2 : the submitted quality such that $P_a(p_2) = 0.10$ (also called LQL);

p^* : maximum allowable percent defective (MAPD);

h^* : relative slope at ' p^* ';

n_1 : sample size of variable sampling plan;

n_2 : sample size of attribute sampling plan;

c_N : acceptance number of normal inspection;

c_T : acceptance number of tightened inspection;

β_j : probability of acceptance for lot quality ' p_j ';

β_j' : probability of acceptance assigned to first stage for percent defective ' p_j ';

β_j'' : probability of acceptance assigned to second stage for percent defective ' p_j ';

d : observed number of nonconforming units in a sample of n units;

$z(j)$: ' z ' value for the j^{th} ordered observation; and

k : variable factor such that a lot is accepted if $\bar{X} \leq A = U - k\sigma$.

Operating Procedure of Mixed Sampling Plan with QSS-1($n;c_N,c_T$) as Attribute Plan

Schilling (1967) provided the following procedure for the independent mixed sampling plan with the upper specification limit (U) and known standard deviation (σ).

- Determine the parameters of the mixed sampling plan n_1, n_2, k, c_N and c_T .
- Select a random sample of size n_1 from the lot.
- If a sample average $\bar{X} \leq A = U - k\sigma$, accept the lot.
- If a sample average $\bar{X} > A = U - k\sigma$, go to step 1.

Step 1: From a lot, take a random sample of size n_2 at the normal level. Count the number of defectives, d :

- If $d \leq c_N$, accept the lot and repeat step 1;

- If $d > c_N$, reject the lot and go to step 2.

Step 2: From the next lot, take a random sample of size n_2 at the tightened level. Count the number of defectives, d :

- If $d \leq c_T$, accept the lot and use step 1 for the next lot;
- If $d > c_T$, reject the lot and repeat step 2 for the next lot.

Construction of Mixed Sampling Plan having QSS-1($n; c_N, c_T$) as Attribute Plan

The operation of mixed sampling plans can be properly assessed by the OC curve for given values of the fraction defective. The development of mixed sampling plans and the subsequent discussions are limited only to the upper specification limit, U . A parallel discussion can be made for lower specification limits.

The procedure for the construction of mixed sampling plans is provided by Schilling (1967) for a given n_1 and a point p_j on the OC curve is:

- Assume that the mixed sampling plan is independent.
- Split the probability of acceptance (β_j) determining the probability of acceptance that will be assigned to the first stage, term this β_j' .
- Determine the sample size n_1 (for variable sampling plan) to be used.
- Calculate the acceptance limit for the variable sampling plan as:

$$A = U - k\sigma = U - [z(p_j) + \{z(\beta_j') / \sqrt{n_1}\}] \sigma$$

where U is the upper specification limit and $z(t)$ is the standard normal variate corresponding to t such that

$$t = \int_{z(t)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du.$$

- Determine the sample average, $\bar{X}$. If a sample average $\bar{X} > A = U - k\sigma$, take a second stage sample of size n_2 using attribute sampling plan.
- Split the probability of acceptance β_j as β_j' and β_j'' , such that $\beta_j = \beta_j' + (1 - \beta_j')\beta_j''$. Fix the value of β_j' .
- Determine β_j'' , the probability of acceptance assigned to the attributes plan associated with the second stage sample as $\beta_j'' = (\beta_j - \beta_j') / (1 - \beta_j')$.
- Determine the appropriate second stage sample of size n_2 from $Pa(p) = \beta_j''$ for $p = p_j$.

Using this procedure, tables can be constructed to facilitate selection of an appropriate mixed sampling plan with QSS-1($n; c_N, c_T$) plan as an attribute plan indexed through MAPD and LQL.

According to Soundararajan and Arumainayagam (1988), the operating characteristic function of QSS-1 is:

$$P_a(p) = \frac{b}{1 - a + b} \quad (1)$$

where

$$a = \sum_{i=0}^{c_N} \frac{e^{-n_2 p} (n_2 p)^i}{i!} \quad (2)$$

and

$$b = \sum_{j=0}^{c_T} \frac{e^{-n_2 p} (n_2 p)^j}{j!} \quad (3)$$

(for acceptance number tightening).

Construction of Sampling Plans Indexed Through MAPD

MAPD (p_*), introduced by Mayer (1967) and further studied by Soundararajan (1975), is the quality level corresponding to the inflection point of the OC curve. The degree of sharpness of inspection about this quality level p_* is measured by p_t , the point at which the tangent to the OC curve at the inflection point cuts the proportion defective axis. For designing

a mixed sampling plan, Soundararajan (1975) proposed a selection procedure indexed with MAPD and $K = p_t/p_*$.

Using the probability mass function of QSS-1 (see expression (1)), the inflection point (p_*) is obtained using

$$\frac{d^2 p_a(p)}{dp^2} = 0$$

and

$$\frac{d^3 p_a(p)}{dp^3} \neq 0.$$

The relative slope of the OC curve is

$$h_* = \left[\frac{-p}{p_a(p)} \right] \frac{dp_a(p)}{dp}$$

at $p = p_*$. The inflection tangent of the OC curve cuts the p axis at $p_t = p_* + (p_*/h_*)$. The values of n_2p_* , h_* , n_2p_t and $R = p_t/p_*$ are calculated for different values of c_N and c_T for $\beta_2' = 0.04$ using a c++ program (see Table 1).

Selection of the Plan

For the given values of p_* and p_t , the ratio $R = \frac{p_t}{p_*}$ is found and the nearest value of

R is located in Table 1. The corresponding value of c_N , c_T and np_* values are noted and the value of n_2 is obtained using $n_2 = \frac{n_2 p_*}{p_*}$.

Example 1

Given $p_* = 0.037$, $p_t = 0.051$ and $\beta_2' = 0.04$, the ratio $R = \frac{p_t}{p_*} = 1.3784$. As shown in

Table 1, the nearest R value is 1.3791 which corresponds to $c_N = 5$ and $c_T = 1$. The value $n_2p_* = 3.3452$ is found, hence the value of n_2 is determined to be $n_2 = \frac{n_2 p_*}{p_*} = \frac{3.3452}{0.037} = 90$.

Thus $n_2 = 90$, $c_N = 5$ and $c_T = 1$ are the parameters selected for the mixed sampling plan having QSS-1($n; c_N, c_T$) as an attribute plan using

the Poisson distribution as a baseline distribution for the given values of $p_* = 0.037$ and $p_t = 0.051$. Construction of Mixed Sampling Plan indexed through LQL

The described procedure is used to construct the mixed sampling plan indexed through LQL(p_2). Assuming the probability of acceptance of the lot be $\beta_2 = 0.10$ and $\beta_2' = 0.04$, the n_2p_2 values are calculated for different values of c_N and c_T using a c++ program (see Table 1).

Selection of the Plan

Table 1 is used to construct the plans when LQL(p_2), c_N and c_T are given. For any given values of p_2 , c_N and c_T one can determine n_2 value using $n_2 = \frac{n_2 p_2}{p_2}$.

Example 2

Given $p_2 = 0.06$, $c_N = 3$ and $c_T = 1$ and $\beta_2' = 0.04$. Using Table 1, find $n_2 = \frac{n_2 p_2}{p_2} = \frac{4.8136}{0.06} = 80$. Thus $n_2 = 80$, $c_N = 3$ and

$c_T = 1$ are the parameters selected for the mixed sampling plan having QSS-1($n; c_N, c_T$) as attribute plan for a specified $p_2 = 0.06$, $c_N = 3$ and $c_T = 1$.

Selection of the Plan

Table 1 is used to construct the plans when LQL (p_2), ' c_N ' and ' c_T ' are given. For any given values of p_2 , c_N and c_T one can determine n_2 value using $n_2 = \frac{n_2 p_2}{p_2}$.

Comparison of Mixed Sampling Plan Indexed through MAPD and LQL

By fixing parameters c_N , c_T and β_j' for specified values of p_* and p_t and assuming $\beta_2' = 0.04$, the values of c_N , c_T and n_2 indexed through MAPD can be determined. Fixing the values of c_N and c_T , the value of p_2 is found by equating $Pa(p) = \beta_2 = 0.10$. Using $\beta_2' = 0.04$, c_N and c_T the value of n_2 is determined using $n_2 = \frac{n_2 p_2}{p_2}$

from Table 1. Using different combinations of p_* , p_t , c_N and c_T , the values of n_2 (indexed through MAPD) and n_2 (indexed through LQL) calculated are presented in Table 2.

Table 1: Various Characteristics of the Mixed Sampling Plan when $\beta_*' = \beta_2' = 0.04$ and $\beta_2 = 0.10$

c_N	c_T	n_2p_2	β_*''	n_2p_*	h_*	n_2p_t	$R = p_t/p_*$
1	0	2.9412	0.6690	0.8538	0.7703	1.9622	2.2982
2	1	4.6086	0.6261	1.8158	1.1034	3.4614	1.9063
2	0	3.1894	0.6095	1.3787	1.3370	2.4099	1.7480
3	2	6.1101	0.6044	2.7979	1.3622	4.8519	1.7341
3	1	4.8136	0.5897	2.3815	1.6231	3.8488	1.6161
3	0	3.4844	0.5811	1.8374	1.8735	2.8181	1.5337
6	5	10.2286	0.5734	5.7797	1.9506	8.7427	1.5127
5	3	7.6959	0.5692	4.3733	2.0587	6.4976	1.4857
7	6	11.5328	0.5670	6.7811	2.1146	9.9879	1.4729
6	4	9.0528	0.5628	5.3702	2.2404	7.7672	1.4464
4	0	3.8063	0.5630	2.2695	2.4041	3.2135	1.4160
6	3	7.9097	0.5556	4.9005	2.5342	6.8342	1.3946
5	1	5.3626	0.5554	3.3452	2.6376	4.6135	1.3791
9	7	12.9515	0.5500	8.3665	2.7059	11.4584	1.3696
6	2	6.7933	0.5500	4.3779	2.8356	5.9218	1.3527
5	0	4.1442	0.5508	2.6858	2.9293	3.6027	1.3414
9	6	11.8549	0.5443	7.9073	3.0102	10.5341	1.3322
6	1	5.6790	0.5455	3.7913	3.1439	4.9972	1.3181
8	4	9.4914	0.5425	6.4072	3.1740	8.4259	1.3151
9	5	10.7925	0.5396	7.4154	3.3249	9.6457	1.3008
8	3	8.4444	0.5384	5.8715	3.4952	7.5514	1.2861
7	1	6.0117	0.5379	4.2248	3.6519	5.3817	1.2738
8	2	7.4095	0.5350	5.2914	3.8241	6.6751	1.2615
7	0	4.8464	0.5349	3.4926	3.9742	4.3714	1.2516
8	1	6.3558	0.5329	4.6471	4.1519	5.7664	1.2409
9	2	7.7425	0.5297	5.7313	4.3226	7.0572	1.2313
9	1	6.7080	0.5272	5.0664	4.6691	6.1515	1.2142
9	0	5.5669	0.5243	4.2808	5.0222	5.1332	1.1991
11	2	8.4386	0.5218	6.5894	5.3246	7.8269	1.1878
11	1	7.4283	0.5453	5.8361	5.3854	6.9198	1.1857
12	0	6.6634	0.5217	5.4307	6.4898	6.2675	1.1541

Table 2: Comparison of Plans Indexed Through MAPD and LQL

p^*	p_t	c_N	c_T	Indexed Through	
				MAPD	LQL
				n_2	n_2
0.024*	0.030	7	0	146	155
0.042	0.058	6	3	117	126
0.045	0.062	5	1	74	80
0.048	0.068	4	0	47	51

*OC curves are drawn.

Figure 1: OC Curves for QSS-1(146;7,0) and (155;7,0)

OC Curve Construction

The OC curves for the plans $n_2 = 146$, $c_N = 7$, $c_T = 0$ (indexed through MAPD) and $n_2 = 155$, $c_N = 7$, $c_T = 0$ (indexed through LQL) based on the different values of $n_2 p_2$ and $p_a(p)$ are presented in Figure 1.

Conclusion

This article used the operating procedure of a mixed sampling plan with QSS-1($n; c_N, c_T$) as an attribute plan and constructed tables for a mixed sampling plan indexed through parameters MAPD and LQL using the Poisson distribution

as a baseline. It may be concluded based on study results that the second stage sample size required for a QSS-1(n ; c_N , c_T) plan indexed through MAPD is less than that of a second stage sample size of the QSS-1(n ; c_N , c_T) plan indexed through LQL. Examples were provided for a specified value of $\beta_j' = 0.04$. If engineers know the levels of MAPD or LQL they can reference the tables provided to determine their sampling plans on site at a factory; this provides flexibility to floor engineers in determining appropriate sampling plans. Various plans can also be constructed to make a system user-friendly by changing the first stage probabilities (β_1' , β_2') and can also be compared for their efficiency.

References

- Devaarul, S. (2003). *Certain studies relating to mixed sampling plans and reliability based sampling plans*. Unpublished PhD Dissertation. Bharathiar University, Tamil Nadu, India.
- Dodge, H. F. (1967). *A new dual system of acceptance sampling: Technical report no.16*. New Brunswick, NJ: The Statistics Centre, Rutgers State University.
- Govindaraju, K. (1991). Procedures and tables of the selection of zero acceptance number quick switching system for compliance testing. *Communications in Statistics-Simulation and Computation*, 20(1), 157-172.
- Mayer, P. L. (1967). A note on sum of Poisson probabilities and an application. *Annals of the Institute of Statistical Mathematics*, 19, 537-542.
- Radhakrishnan, R., & Sampath Kumar, R. (2006a). Construction of mixed sampling plan indexed through MAPD and IQL with single sampling plan as attribute plan. *National Journal of Technology*, 2(2), 26-29.
- Radhakrishnan, R., & Sampath Kumar, R. (2006b). Construction of mixed sampling plans indexed through MAPD and AQL with chain sampling plan as attribute plan. *STARS International Journal*, 7(1), 14-22.
- Radhakrishnan, R., & Sampath Kumar, R. (2007a). Construction of mixed sampling plans indexed through MAPD and IQL with double sampling plan as attribute plan. *The Journal of the Kerala Statistical Association*, 18, 13-22.
- Radhakrishnan, R., & Sampath Kumar, R. (2007b). Construction of mixed sampling plans indexed through MAPD and AQL with double sampling plan as attribute plan. *The International Journal of Statistics and Systems*, 2(2), 33-39.
- Radhakrishnan, R., & Sampath Kumar, R. (2009). Construction and comparison of mixed sampling plans having ChSP-(0, 1) plan as attribute plan. *International Journal of Statistics and Management Systems*, 4(1-2), 134-149.
- Radhakrishnan, R., Sampath Kumar, R., & Saravanan, P. G. (2009). Construction of dependent mixed sampling plans using single sampling plan as attribute plan. *International Journal of Statistics and Systems*, 4(1), 67-74.
- Romboski, L. D. (1969). *An investigation of quick switching acceptance sampling system*. Unpublished PhD Dissertation. New Brunswick, NJ, Rutgers State University.
- Sampath Kumar, R. (2007). *Construction and selection of mixed variables-attributes sampling plan*. Unpublished PhD Dissertation. Department of Statistics, Bharathiar University, Coimbatore, Tamil Nadu, India.
- Schilling, E. G. (1967). A general method for determining the operating characteristics of mixed variables – attributes sampling plans single side specifications, S. D. known. Unpublished PhD Dissertation, New Brunswick, NJ, Rutgers State University.
- Soundarajan, V. (1975). Maximum allowable percent defective (MAPD) single sampling inspection by attribute plan. *Journal of Quality Technology*, 7(4), 173-177.
- Soundararajan, V., & Arumainayagam, S. D. (1988). *Modifications of quick switching systems with special reference to crossover point of their composite OC curves: Research report no.25*. Bharathiar University, Coimbatore: Department of Statistics.

On Some Negative Integer Moments of Quasi-Negative-Binomial Distribution

Anwar Hassan
King Saud University
Riyadh, Kingdom of Saudi Arabia

Sheikh Bilal
A. S. College, Srinagar
Kashmir, India

Negative integer moments of the quasi-negative-binomial distribution (QNBD) are investigated. This distribution includes recurrence relations which are helpful in the solution of many applied statistical problems, particularly in life testing and survey sampling, where ratio estimators are useful. Results study show the negative-binomial distribution when the parameter θ_2 is zero and also indicate the mean of the QNBD model when its parameters are changed.

Key words: Quasi-negative-binomial distribution, recurrence relations, Abel series expansion, negative-binomial distribution.

Introduction

The quasi-negative-binomial distribution (QNBD) was introduced in different forms by Janardan (1975), Nandi and Das (1994) and Sen and Jain (1996) but has not been studied in detail. The discrete probability function of the QNBD is given by

$$P_x(a, \theta_1, \theta_2) = \frac{(a+x-1)!}{(a-1)! x!} \frac{\theta_1(\theta_1 + x\theta_2)^{x-1}}{(1+\theta_1 + x\theta_2)^{x+a}},$$

$$x = 0, 1, 2, \dots \quad (1.1)$$

$$(\theta_1 + x\theta_2) \geq 0; a > 0, \theta_1 > 0 \text{ if } \theta_2 > 0$$

and

$$(\theta_1 + x\theta_2) \geq 0; \text{ if } \theta_2 < 0$$

where (a, θ_1, θ_2) are parameters of the distribution. When θ_2 is negative, the

probabilities of the QNBD model become negative. In addition, there appears to be a natural truncation for x , for which $P_x(a, \theta_1, \theta_2) = 0$; however, this has not been verified and requires a detailed error analysis, which is not included herein.

The QNBD model reduces to a negative-binomial distribution (NBD) model at $\theta_2 = 0$. It appears from the model that the β parameter in Greenwood and Yule's (1920) NBD model was replaced by $(\theta_1 + x\theta_2)$, where x is the number of occurrences; this implies that, with successive occurrences, there is some changing tendency in the θ_1 parameter.

Hassan and Bilal (2008) explored the properties of the QNBD model (1.1) with mean and variance obtained in a hypergeometric function given as

$$\mu'_1 = a\theta_1 {}_2F_0[1, a+1, _; \theta_2] \quad (1.2)$$

$$\begin{aligned} \mu'_2 = & a\theta_1 {}_2F_0[1, a+1, _; \theta_2] \\ & + \theta_1(\theta_1 + 2\theta_2)a(a+1) {}_2F_0[2, a+2, _; \theta_2] \\ & + \theta_1\theta_2^2 a(a+1)(a+2) {}_2F_0[3, a+3, _; \theta_2] \end{aligned} \quad (1.3)$$

where ${}_2F_0[1, a+1, _; \theta_2]$ is a hypergeometric function defined by

Anwar Hassan is an Associate Professor in the Department of Statistics & Operations Research. Email him at: anwar.hassan2007@gmail.com or amohammadabdulkhair@ksu.edu.sa. Sheikh Bilal is a Senior Assistant Professor in the Department of Statistics. Email him at sbilal_sbilal@yahoo.com.

$${}_2F_0[1, a+1, _ ; \theta_2] = \sum_{j=0}^{\infty} 1^{[j]} (a+1)^{[j]} \frac{\theta_2^j}{j!}.$$

Hassan and Bilal (2006) found applications for the QNBD model in queuing theory, theories of microorganisms and biology. They investigated the distribution of numbers of accidents as a QNBD model using Irwin's (1968) theory of proneness-liability model and then applied the model to hunting accidents, home injuries and strikes in industries; they obtained better model fits than Consul and Jain's (1973), using a generalized Poisson distribution (GPD) model.

A difficulty with the QNBD model is that its moments appear in an infinite series, which does not seem to converge to an expression that will produce moment estimators. This article investigates negative integer moments of the QNBD. This distribution includes recurrence relations which are helpful in the solution of many applied statistics problems. Results from this study show the negative-binomial distribution when the parameter θ_2 is zero and indicate the mean of the QNBD model when its parameters are changed.

Negative Integer Moments

Suppose that $\phi_s(k, a, \theta_1) = E[x+k]^{-s}$ denotes the s^{th} negative integer moments of the QNBD model (1.1), then the following results on the negative integer moments are true for the proposed model:

$$\begin{aligned} \varphi_1\left(\frac{\theta_1}{\theta_2}, a, \theta_1\right) &= E\left(x + \frac{\theta_1}{\theta_2}\right)^{-1} \\ &= \frac{(\theta_1 + \theta_2 - a\theta_1\theta_2)\theta_2}{\theta_1(\theta_1 + \theta_2)} \end{aligned} \quad (1.4)$$

$$\begin{aligned} \varphi_2\left(\frac{\theta_1}{\theta_2}, a, \theta_1\right) &= E\left(x + \frac{\theta_1}{\theta_2}\right)^{-2} \\ &= \frac{\theta_2}{\theta_1} \left(\frac{\theta_2}{\theta_1} - \frac{a\theta_2^2}{(\theta_1 + \theta_2)} \right) \\ &\quad + \frac{a\theta_2^2}{(\theta_1 + \theta_2)} \times \\ &\quad \left(\frac{\theta_2}{(\theta_1 + \theta_2)} - \frac{(a+1)\theta_2^2}{(\theta_1 + 2\theta_2)} \right) \end{aligned} \quad (1.5)$$

$$\begin{aligned} \varphi_s\left(\frac{\theta_1}{\theta_2}, a, \theta_1\right) &= \frac{\theta_2}{\theta_1} \varphi_{s-1}\left(\frac{\theta_1}{\theta_2}, a, \theta_1\right) - \\ &\quad \frac{a\theta_2^2}{(\theta_1 + \theta_2)} \varphi_{s-1} \times \\ &\quad \left(\frac{\theta_1 + \theta_2}{\theta_2}, a+1, \theta_1 + \theta_2 \right) \end{aligned} \quad (1.6)$$

$$\begin{aligned} \varphi_1\left(\frac{1+\theta_1}{\theta_2}, a, \theta_1\right) &= E\left(x + \frac{1+\theta_1}{\theta_2}\right)^{-1} \\ &= \frac{\theta_2(a\theta_2 - 1)}{(a\theta_2 - \theta_1 - 1)} \end{aligned} \quad (1.7)$$

$$\begin{aligned} \varphi_2\left(\frac{1+\theta_1}{\theta_2}, a, \theta_1\right) &= E\left(x + \frac{1+\theta_1}{\theta_2}\right)^{-2} \\ &= \frac{\theta_2^2(1 - a\theta_2)}{(1 + \theta_1 - a\theta_2)^2} \\ &\quad - \frac{a\theta_2^3(1 - (a+1)\theta_2)}{(1 + \theta_1 - (a+1)\theta_2)(1 + \theta_1 - a\theta_2)} \end{aligned} \quad (1.8)$$

$$\begin{aligned} \varphi_s \left(\frac{1+\theta_1}{\theta_2}, a, \theta_1 \right) &= \frac{\theta_2}{(1+\theta_1-a\theta_2)} \varphi_{s-1} \times \\ &\quad \left(\frac{1+\theta_1}{\theta_2}, a, \theta_1 \right) - \\ &\quad \frac{a\theta_2^2}{(1+\theta_1-a\theta_2)} \varphi_{s-1} \times \\ &\quad \left(\frac{1+\theta_1}{\theta_2}, a+1, \theta_1 \right) \end{aligned} \quad (1.9)$$

$$\phi_1(a, a+1, \theta_1) = \frac{(1-a\theta_2)}{a(1+\theta_1-a\theta_2)} \quad (1.10)$$

$$\phi_1(1, a+1, \theta_1+\theta_2) = \frac{(\theta_1+\theta_2-a\theta_1\theta_2)}{a\theta_1^2} \quad (1.11)$$

$$\begin{aligned} \varphi_1(k, a, \theta_1) &= \frac{(a-k-1)}{(a-1)(1+\theta_1-k\theta_2)} \varphi_1(k, a-1, \theta_1) \\ &\quad + \frac{1-\theta_2(a-1)}{(1+\theta_1-k\theta_2)(a-1)} \end{aligned} \quad (1.12)$$

$$\begin{aligned} \varphi_s(0, a, \theta_1) &= \frac{a\theta_1^2}{(\theta_1+\theta_2)} \varphi_{s+1}(1, a+1, \theta_1+\theta_2) \\ &\quad + \frac{a\theta_1\theta_2}{(\theta_1+\theta_2)} \varphi_s(1, a+1, \theta_1+\theta_2) \end{aligned} \quad (1.13)$$

Proof

Taking the summation of (1.1) and differentiating it with respect to θ_1 , results in the simplification

$$\frac{1}{\theta_1} + \sum_{x=0}^{\infty} \frac{(x-1)}{(\theta_1+x\theta_2)} \frac{(a+x-1)!}{(a-1)! x!} \frac{\theta_1(\theta_1+x\theta_2)^{x-1}}{(1+\theta_1+x\theta_2)^{a+x}} = a, \quad (1.14)$$

and writing $\theta_2(x-1) = (\theta_1+x\theta_2) - (\theta_1+\theta_2)$, results in

$$\frac{(\theta_1+\theta_2)}{\theta_1\theta_2} - \frac{(\theta_1+\theta_2)}{\theta_2^2} E \left(x + \frac{\theta_1}{\theta_2} \right)^{-1} = a.$$

After rearranging the terms in the equation result (1.4) follows. Similarly, result (1.5) can be obtained by differentiating (1.14) with respect to θ_1 and simplifying the resulting equation. The results represented by (1.4) and (1.5) can also be obtained from recurrence relation (1.6), which is proven by:

$$\begin{aligned} &\varphi_s \left(\frac{\theta_1}{\theta_2}, a, \theta_1 \right) \\ &= E \left(x + \frac{\theta_1}{\theta_2} \right)^{-s} \\ &= \sum_{x=0}^{\infty} \left(x + \frac{\theta_1}{\theta_2} \right)^{-s} \frac{(a+x-1)!}{(a-1)! x!} \frac{\theta_1(\theta_1+x\theta_2)^{x-1}}{(1+\theta_1+x\theta_2)^{a+x}} \\ &= \left[\frac{\theta_2}{\theta_1} \sum_{x=0}^{\infty} \left(x + \frac{\theta_1}{\theta_2} \right)^{-s+1} \right. \\ &\quad \left. \times \left(1 - \frac{x\theta_2}{(\theta_1+x\theta_2)} \right) \frac{(a+x-1)!}{(a-1)! x!} \frac{\theta_1(\theta_1+x\theta_2)^{x-1}}{(1+\theta_1+x\theta_2)^{a+x}} \right] \\ &= \left[\frac{\theta_2}{\theta_1} \sum_{x=0}^{\infty} \left(x + \frac{\theta_1}{\theta_2} \right)^{-s+1} \right. \\ &\quad \left. P_x(a, \theta_1, \theta_2) - a\theta_2^2 \sum_{x=1}^{\infty} \left(x + \frac{\theta_1}{\theta_2} \right)^{-s+1} \frac{(a+x-1)!}{a! (x-1)!} \frac{(\theta_1+x\theta_2)^{x-2}}{(1+\theta_1+x\theta_2)^{a+x}} \right] \end{aligned}$$

where $P_x(a, \theta_1, \theta_2)$ is defined by (1.1). Replacing x with $(x+1)$ in the second component of the equation results in:

$$\begin{aligned} \varphi_s\left(\frac{\theta_1}{\theta_2}, a, \theta_1\right) &= \\ \frac{\theta_2}{\theta_1} \varphi_{s-1}\left(\frac{\theta_1}{\theta_2}, a, \theta_1\right) &- \\ \frac{a\theta_2^2}{(\theta_1 + \theta_2)} \sum_{x=0}^{\infty} \left(x + \frac{(\theta_1 + \theta_2)}{\theta_2}\right) P_x(a+1, \theta_1 + \theta_2, \theta_2), \end{aligned}$$

which gives recurrence relation (1.6). To prove result (1.7), take the summation of QNBD model (1.1) with parameters $(a+1, \theta_1, \theta_2)$ to yield:

$$\sum_{x=0}^{\infty} \frac{(a+x)!}{a!x!} \frac{\theta_1(\theta_1 + x\theta_2)^{x-1}}{(1 + \theta_1 + x\theta_2)^{a+1+x}} = 1.$$

Rewriting this equation as

$$\frac{1}{a} \sum_{x=0}^{\infty} \frac{(a+x)}{(1 + \theta_1 + x\theta_2)} P_x(a, \theta_1, \theta_2) = 1,$$

and writing $\theta_2(a+x)$ as a sum of two components $(1 + \theta_1 + x\theta_2)$ and $(a\theta_2 - \theta_1 - 1)$, results in

$$\frac{(a\theta_2 - \theta_1 - 1)}{a\theta_2^2} \varphi_1\left(\frac{1 + \theta_1}{\theta_2}, a, \theta_1\right) + \frac{1}{a\theta_2} = 1.$$

Rearranging the terms in the equation result (1.7) follows. Taking the summation of the QNBD model with parameters $(a+2, \theta_1, \theta_2)$ and proceeding in the same way, result (1.6) is obtained. To prove recurrence relation (1.9):

$$\begin{aligned} \varphi_s\left(\frac{1 + \theta_1}{\theta_2}, a, \theta_1\right) &= \\ = \sum_{x=0}^{\infty} \left(x + \frac{1 + \theta_1}{\theta_2}\right)^{-s+1} &\left[\frac{\left(\frac{\theta_2}{(1 + \theta_1 + x\theta_2)}\right)}{\times \left(\frac{(a+x-1)!}{(a-1)!x!} \frac{\theta_1(\theta_1 + x\theta_2)^{x-1}}{(1 + \theta_1 + x\theta_2)^{a+x}}\right)} \right] \\ = \theta_2 \sum_{x=0}^{\infty} \left(x + \frac{1 + \theta_1}{\theta_2}\right)^{-s+1} &\left[\frac{\left(1 - \frac{\theta_1 + x\theta_2}{(1 + \theta_1 + x\theta_2)}\right)}{\times \left(\frac{(a+x-1)!}{(a-1)!x!} \frac{\theta_1(\theta_1 + x\theta_2)^{x-1}}{(1 + \theta_1 + x\theta_2)^{a+x}}\right)} \right] \end{aligned}$$

and writing $(\theta_1 + x\theta_2) = (\theta_1 - a\theta_2) + (a+x)\theta_2$, results in the simplification:

$$\begin{aligned} \varphi_s\left(\frac{1 + \theta_1}{\theta_2}, a, \theta_1\right) &= \\ \theta_2 \varphi_{s-1}\left(\frac{1 + \theta_1}{\theta_2}, a, \theta_1\right) &- \\ (\theta_1 - a\theta_2) \varphi_s\left(\frac{1 + \theta_1}{\theta_2}, a, \theta_1\right) &- \\ a\theta_2^2 \varphi_{s-1}\left(\frac{1 + \theta_1}{\theta_2}, a+1, \theta_1\right). \end{aligned}$$

After rearranging the terms in the equation result (1.9) follows. The results (1.10), (1.11) and (1.12) are straightforward and can be obtained in a similar way, however, for recurrence relation (1.13):

$$\begin{aligned} \varphi_s(0, a, \theta_1) &= \sum_{x=0}^{\infty} x^{-s} \frac{(a+x-1)!}{(a-1)!x!} \frac{\theta_1(\theta_1 + x\theta_2)^{x-1}}{(1 + \theta_1 + x\theta_2)^{a+x}} \\ &= \sum_{x=1}^{\infty} x^{-s-1} \frac{(a+x-1)!}{(a-1)!x!} \frac{\theta_1(\theta_1 + x\theta_2)^{x-1}}{(1 + \theta_1 + x\theta_2)^{a+x}} \end{aligned}$$

and replacing x by $(x+1)$ in the equation above results in

$$\varphi_s(0, a, \theta_1) =$$

$$\frac{a\theta_1}{(\theta_1 + \theta_2)} \times$$

$$\sum_{x=0}^{\infty} (x+1)^{-s-1} (\theta_1 + \theta_2 (x+1)) P_x(a+1, \theta_1 + \theta_2, \theta_2)$$

which gives

$$\varphi_s(0, a, \theta_1) =$$

$$\frac{a\theta_1^2}{(\theta_1 + \theta_2)} \varphi_{s+1}(1, a+1, \theta_1 + \theta_2) +$$

$$\frac{a\theta_1\theta_2}{(\theta_1 + \theta_2)} \varphi_s(1, a+1, \theta_1 + \theta_2)$$

Charalambides (1990) examined an extension of the class of power series distributions and obtained a discrete class of Abel series distributions. He also explored its properties with an application to the fluctuations of sample function of stochastic process. Nandi and Das (1994) also obtained a class of Abel series distributions. Hassan and Bilal (2008) showed that the QNBD model belongs to a family of Abel series distributions by taking the Abel series expansion of $(c-r)^{-a}$ given as

$$(c-r)^{-a} = \sum_{x=0}^{\infty} \frac{(a+x-1)! r(r+bx)^{x-1}}{(a-1)! x! (c+bx)^{a+x}}, \quad (1.15)$$

where

$$\frac{r}{(c-r)} = \theta_1, \frac{b}{(c-r)} = \theta_2$$

and

$$\frac{c}{(c-r)} = 1 + \frac{r}{(c-r)} = 1 + \theta_1.$$

The expression in (1.15) gives the sum of the QNBD model, which is equal to unity. The following results, obtained on the basis of (1.15), are proven as:

$$\phi_1(1, a, \theta_1) = E(x+1)^{-1} = \frac{\theta_1}{(a-1)(\theta_1 - \theta_2)^2} - \frac{\theta_2}{(\theta_1 - \theta_2)} \quad (1.16)$$

$$\begin{aligned} \varphi_1(2, a, \theta_1) &= E(x+1)^{-2} \\ &= \frac{\theta_1}{(a-1)(\theta_1 - 2\theta_2)^2} - \frac{\theta_1}{(a-1)(a-2)(\theta_1 - 2\theta_2)^3} \\ &\quad - \frac{\theta_2}{(\theta_1 - 2\theta_2)} \end{aligned} \quad (1.17)$$

$$\begin{aligned} \varphi_1[(1, 2), a, \theta_1] &= E[(x+1)(x+2)]^{-1} \\ &= \frac{\theta_1\theta_2(3\theta_2 - 2\theta_1)}{(a-1)(\theta_1 - \theta_2)^2(\theta_1 - 2\theta_2)^2} \\ &\quad + \frac{\theta_1}{(a-1)(a-2)(\theta_1 - 2\theta_2)^3} \\ &\quad + \frac{\theta_2}{(\theta_1 - \theta_2)(\theta_1 - 2\theta_2)} \end{aligned} \quad (1.18)$$

$$\begin{aligned} \varphi_1(a-1, a, \theta_1) &= E(x+a-1)^{-1} \\ &= \frac{1 - \theta_2(a-1)}{(a-1)[1 + \theta_1 - \theta_2(a-1)]} \end{aligned} \quad (1.19)$$

$$\begin{aligned} \varphi_1(a-2, a, \theta_1) &= E(x+a-2)^{-1} \\ &= \frac{[1 - \theta_2(a-1)][1 + \theta_1 - \theta_2(a-1)]}{(a-1)(a-2)[(1 + \theta_1 - \theta_2(a-1))^2 + 2\theta_2(1 + \theta_1 - \theta_2(a-1)) + \theta_2^2]} \end{aligned} \quad (1.20)$$

Proof

Integrating (1.15) with respect to r , results in

$$\begin{aligned}
 & \frac{(c-r)^{-a+1}}{(a-1)} \\
 &= \sum_{x=0}^{\infty} \frac{(a+x-1)!}{(a-1)! x!} \left(\frac{\frac{r(r+bx)^x}{x}}{-\frac{(r+bx)^{x+1}}{x(x+1)}} \right) (c+bx)^{-a-x} \\
 &= \sum_{x=0}^{\infty} \frac{(a+x-1)!}{(a-1)! x!} \left(\frac{\frac{(r+bx)}{x}}{-\frac{(r+bx)^2}{x(x+1)r}} \right) \frac{r(r+bx)^{x-1}}{(c+bx)^{a+x}} \\
 & \quad (1.21)
 \end{aligned}$$

Expressing the equation in terms of θ_1 and θ_2 results in

$$\frac{1}{(a-1)} = \sum_{x=0}^{\infty} \left(\frac{(\theta_1 + \theta_2 x)}{x} - \frac{(\theta_1 + \theta_2 x)^2}{x(x+1)\theta_1} \right) P_x(a, \theta_1, \theta_2)$$

and writing

$$\begin{aligned}
 & (\theta_1 + x\theta_2)^2 \\
 &= (\theta_1 - \theta_2)^2 + 2\theta_2(\theta_1 - \theta_2)(x+1) + \theta_2^2(x+1)^2,
 \end{aligned}$$

the result (1.16) follows based on simplifications. Again, integrating (1.21) with respect to r , result (1.17) is obtained. Result (1.15) follows from (1.13) and (1.14) by using the relation

$$E[(x+1)(x+2)]^{-1} = E(x+1)^{-1} - E(x+2)^{-1}.$$

To prove result (1.19), integrating (1.15) with respect to c , results in

$$\frac{(c-r)^{-a+1}}{(a-1)} = \sum_{x=0}^{\infty} \frac{(c+bx)}{(x+a-1)} \frac{(a+x-1)!}{(a-1)! x!} \frac{r(r+bx)^{x-1}}{(c+bx)^{a+x}}. \quad (1.22)$$

Expressing the equation above in terms of θ_1 and θ_2 results in

$$\frac{1}{(a-1)} = \sum_{x=0}^{\infty} \frac{(1+\theta_1+x\theta_2)}{(x+a-1)} P_x(a, \theta_1, \theta_2).$$

Writing

$$(1+\theta_1+x\theta_2) = [1+\theta_1-\theta_2(a-1)] + \theta_2(x+a-1),$$

results in

$$\begin{aligned}
 & \frac{1}{(a-1)} \\
 &= \sum_{x=0}^{\infty} \frac{[1+\theta_1-\theta_2(a-1)] + \theta_2(x+a-1)}{(x+a-1)} P_x(a, \theta_1, \theta_2) \\
 &= [1+\theta_1-\theta_2(a-1)]\phi_1(a-1, a, \theta_1) + \theta_2
 \end{aligned}$$

Rearranging the terms in the equation result (1.19) follows. Integrating (1.22) with respect to c and proceeding in the same way results in (1.20).

Another useful set of recurrence relations on the negative integer moments of the QNBD model from which a number of important results can be deduced are

$$\begin{aligned}
 & \varphi_s(k, a+1, \theta_1) \\
 &= \frac{1}{(1+\theta_1-\theta_2 k)} \left(\frac{(a-k)}{a} \varphi_s(k, a, \theta_1) \right. \\
 & \quad \left. + \frac{1}{a} \varphi_{s-1}(k, a, \theta_1) \right. \\
 & \quad \left. - \theta_2 \varphi_{s-1}(k, a+1, \theta_1) \right) \\
 & \quad (1.23)
 \end{aligned}$$

$$\begin{aligned}
 & \varphi_s(k, a, \theta_1) \\
 &= \int_{k\theta_2}^{\theta_1} (\theta_1 - k\theta_2)^k (1+\theta_1 - k\theta_2)^{a-k} [G(a, \theta_1) + \\
 & \quad G(a+1, \theta_1)] d\theta_1 \\
 & \quad (1.24)
 \end{aligned}$$

where $G(a, \theta_1)$ and $G(a+1, \theta_1)$ are defined as

$$G(a, \theta_1) = \left[\left(\frac{\theta_1 + \theta_2}{\theta_1^2} - \frac{(\theta_1 - k\theta_2)}{\theta_1(1 + \theta_1 - k\theta_2)} \right) \varphi_{s-1}(k, a, \theta_1) - \frac{\theta_2}{\theta_1} \varphi'_{s-1}(k, a, \theta_1) \right] \quad (1.25)$$

$$G(a+1, \theta_1) = \frac{a\theta_2}{\theta_1(1 + \theta_1 - k\theta_2)} \varphi_{s-1}(k, a+1, \theta_1) \quad (1.26)$$

Proof

First, result (1.23) is proven, which is subsequently required in the derivation of (1.24). Writing

$$\varphi_s(k, a+1, \theta_1) = \sum_{x=0}^{\infty} (x+k)^{-s} \frac{(a+x)!}{a!x!} \frac{\theta_1(\theta_1 + x\theta_2)^{x-1}}{(1 + \theta_1 + x\theta_2)^{a+1+x}}$$

simplifies to

$$\begin{aligned} \varphi_s(k, a+1, \theta_1) &= \frac{1}{a} \sum_{x=0}^{\infty} (x+k)^{-s} \frac{(a+x)}{(1 + \theta_1 + x\theta_2)} P_x(a, \theta_1, \theta_2) \\ &= \frac{1}{a} \sum_{x=0}^{\infty} (x+k)^{-s} (a+x) \left(1 - \frac{(\theta_1 + x\theta_2)}{(1 + \theta_1 + x\theta_2)} \right) P_x(a, \theta_1, \theta_2) \end{aligned}$$

Writing $(a+x) = (a-k) + (x+k)$ and $(\theta_1 + x\theta_2) = (\theta_1 - k\theta_2) + \theta_2(x+k)$ in the second component of the equation results in

$$\begin{aligned} \varphi_s(k, a+1, \theta_1) &= \frac{(a-k)}{a} \varphi_s(k, a, \theta_1) + \frac{1}{a} \varphi_{s-1}(k, a, \theta_1) \\ &\quad - (\theta_1 - k\theta_2) \varphi_s(k, a+1, \theta_1) - \theta_2 \varphi_{s-1}(k, a+1, \theta_1) \end{aligned}$$

Rearranging the terms, results in (1.23). To prove (1.24), suppose

$$\begin{aligned} U(x) &= \sum_{x=0}^{\infty} (x+k)^{-s} \frac{(a+x-1)!}{(a-1)!x!} \frac{(\theta_1 + x\theta_2)^x}{(1 + \theta_1 + x\theta_2)^{a+x}} \\ &= \frac{1}{\theta_1} \sum_{x=0}^{\infty} (x+k)^{-s} (\theta_1 + x\theta_2) \frac{(a+x-1)!}{(a-1)!x!} \frac{\theta_1(\theta_1 + x\theta_2)^{x-1}}{(1 + \theta_1 + x\theta_2)^{a+x}} \end{aligned} \quad (1.27)$$

Writing

$$(\theta_1 + x\theta_2) = (\theta_1 - k\theta_2) + \theta_2(x+k)$$

results in

$$U(x) = \frac{(\theta_1 - k\theta_2)}{\theta_1} \varphi_s(k, a, \theta_1) + \frac{\theta_2}{\theta_1} \varphi_{s-1}(k, a, \theta_1)$$

and differentiating the equation with respect to θ_1 , results in

$$\begin{aligned} U'(x) &= \frac{(\theta_1 - k\theta_2)}{\theta_1} \varphi'_s(k, a, \theta_1) + \frac{k\theta_2}{\theta_1^2} \varphi_s(k, a, \theta_1) \\ &\quad - \frac{\theta_2}{\theta_1^2} \varphi_{s-1}(a) + \frac{\theta_2}{\theta_1} \varphi'_{s-1}(k, a, \theta_1) \end{aligned} \quad (1.28)$$

where

$$U'(x) = \sum_{x=0}^{\infty} \frac{(x+k)^{-s}}{\theta_1} \left(x - \frac{(a+x)(\theta_1 + x\theta_2)}{(1 + \theta_1 + x\theta_2)} \right) P_x(a, \theta_1, \theta_2)$$

Writing

$$(\theta_1 + x\theta_2) = (\theta_1 - k\theta_2) + \theta_2(x+k)$$

in the equation results in

$U'(x)$

$$\begin{aligned}
 &= \frac{1}{\theta_1} \sum_{x=0}^{\infty} (x+k)^{-s} [(x+k)-k] P_x(a, \theta_1, \theta_2) \\
 &- \frac{a}{\theta_1} \sum_{x=0}^{\infty} (x+k)^{-s} \left[\frac{(\theta_1 - k\theta_2)}{+\theta_2(x+k)} \right] P_x(a+1, \theta_1, \theta_2) \\
 &= \frac{1}{\theta_1} \varphi_{s-1}(k, a, \theta_1) - \frac{k}{\theta_1} \varphi_s(k, a, \theta_1) \\
 &- \frac{a(\theta_1 - k\theta_2)}{\theta_1} \varphi_s(k, a+1, \theta_1) \\
 &- \frac{a\theta_2}{\theta_1} \varphi_{s-1}(k, a+1, \theta_1)
 \end{aligned}$$

Using this equation in (1.28) results in

$$\begin{bmatrix} \frac{1}{\theta_1} \varphi_{s-1}(k, a, \theta_1) \\ -\frac{k}{\theta_1} \varphi_s(k, a, \theta_1) \\ -\frac{a(\theta_1 - k\theta_2)}{\theta_1} \varphi_s(k, a+1, \theta_1) \\ -\frac{a\theta_2}{\theta_1} \varphi_{s-1}(k, a+1, \theta_1) \end{bmatrix} = \begin{bmatrix} \frac{(\theta_1 - k\theta_2)}{\theta_1} \varphi'_s(k, a, \theta_1) \\ +\frac{k\theta_2}{\theta_1^2} \varphi_s(k, a, \theta_1) \\ -\frac{\theta_2}{\theta_1^2} \varphi_{s-1}(k, a, \theta_1) \\ +\frac{\theta_2}{\theta_1} \varphi'_{s-1}(k, a, \theta_1) \end{bmatrix}$$

Rearranging the terms,

$$\begin{bmatrix} \frac{(\theta_1 - k\theta_2)}{\theta_1} \varphi'_s(k, a, \theta_1) \\ +\frac{k(\theta_1 + \theta_2)}{\theta_1^2} \varphi_s(k, a, \theta_1) \end{bmatrix} = \begin{bmatrix} \frac{(\theta_1 + \theta_2)}{\theta_1^2} \varphi_{s-1}(k, a, \theta_1) \\ -\frac{\theta_2}{\theta_1} \varphi'_{s-1}(k, a, \theta_1) \\ -\frac{a(\theta_1 - k\theta_2)}{\theta_1} \varphi_s(k, a+1, \theta_1) \\ -\frac{a\theta_2}{\theta_1} \varphi_{s-1}(k, a+1, \theta_1) \end{bmatrix}$$

and using (1.23) results in

$$\begin{bmatrix} \frac{(\theta_1 - k\theta_2)}{\theta_1} \varphi'_s(k, a, \theta_1) \\ +\frac{k(\theta_1 + \theta_2)}{\theta_1^2} \varphi_s(k, a, \theta_1) \end{bmatrix} = \begin{bmatrix} \frac{(\theta_1 + \theta_2)}{\theta_1^2} \varphi_{s-1}(k, a, \theta_1) \\ -\frac{\theta_2}{\theta_1} \varphi'_{s-1}(k, a, \theta_1) \\ -\frac{a(\theta_1 - k\theta_2)}{\theta_1(1 + \theta_1 - k\theta_2)} \\ \times \left(\frac{(a-k)}{a} \varphi_s(k, a, \theta_1) \right) \\ +\frac{1}{a} \varphi_{s-1}(k, a, \theta_1) \\ -\theta_2 \varphi_{s-1}(k, a+1, \theta_1) \end{bmatrix} - \frac{a\theta_2}{\theta_1} \varphi_{s-1}(k, a+1, \theta_1)$$

Adding similar terms results in

$$\begin{bmatrix} \frac{(\theta_1 - k\theta_2)}{\theta_1} \varphi'_s(k, a, \theta_1) \\ +\left(\frac{k(\theta_1 + \theta_2)}{\theta_1^2} + \frac{(a-k)(\theta_1 - k\theta_2)}{\theta_1(1 + \theta_1 - k\theta_2)} \right) \varphi_s(k, a, \theta_1) \end{bmatrix} = \begin{bmatrix} \varphi'_s(k, a, \theta_1) + \frac{\theta_1}{(\theta_1 - k\theta_2)} \\ \times \left(\frac{k(\theta_1 + \theta_2)}{\theta_1^2} + \frac{(a-k)(\theta_1 - k\theta_2)}{\theta_1(1 + \theta_1 - k\theta_2)} \right) \varphi_s(k, a, \theta_1) \end{bmatrix} = \frac{\theta_1}{(\theta_1 - k\theta_2)} [G(a, \theta_1) - G(a+1, \theta_1)]$$

which gives the linear differential equation

$$\begin{aligned}
 & \varphi'_s(k, a, \theta_1) + \\
 & \frac{\theta_1}{(\theta_1 - k\theta_2)} \left(\frac{k(\theta_1 + \theta_2)}{\theta_1^2} + \frac{(a-k)(\theta_1 - k\theta_2)}{\theta_1(1 + \theta_1 - k\theta_2)} \right) \varphi_s(k, a, \theta_1) \\
 & = \frac{\theta_1}{(\theta_1 - k\theta_2)} [G(a, \theta_1) - G(a+1, \theta_1)] \\
 & \quad (1.29)
 \end{aligned}$$

Where $G(a, \theta_1)$ and $G(a+1, \theta_1)$ are defined in (1.25) and (1.26) respectively. The integrating factor for the differential equation is

$$\begin{aligned}
 I.F. &= \exp \left\{ \int \left(\frac{k(\theta_1 + \theta_2)}{\theta_1(\theta_1 - k\theta_2)} + \frac{(a-k)}{(1 + \theta_1 - k\theta_2)} \right) \partial \theta_1 \right\} \\
 &= \exp \left\{ \int \left(\frac{\theta_1(k+2) - (2\theta_1 - k\theta_2)}{\theta_1(\theta_1 - k\theta_2)} + \frac{(a-k)}{(1 + \theta_1 - k\theta_2)} \right) \partial \theta_1 \right\}.
 \end{aligned}$$

Simplifying this equation gives the integrating factor

$$I.F. = \frac{(\theta_1 - k\theta_2)^{k+1} (1 + \theta_1 - k\theta_2)^{a-k}}{\theta_1}.$$

Multiplying (1.29) with this integrating factor and integrating it with respect to θ_1 from $k\theta_2$ to θ_1 , result (1.24) follows. Note that, taking $\theta_2 = 0$ in (1.24), the recurrence relation for the NBD model is obtained and is given by:

$$\varphi_s(k, a, \theta_1) = \int_0^{\theta_1} \theta_1^{k-1} (1 + \theta_1)^{a-k-1} \varphi_{s-1}(k, a, \theta_1) \partial \theta_1.$$

The mean of QNBD model (1.1) results in an infinite series which renders it useful for estimating parameters by a method of moments. Next, a couple of recurrence relations between

two means when their parameters are changed are proven. Suppose $\mu(a, \theta_1, \theta_2)$ represents the mean of the QNBD model with parameters (a, θ_1, θ_2) , then the ratio of the mean – when the parameter θ_1 is changed to $(\theta_1 + \theta_2)$ – to the mean when parameters are unchanged is independent of parameter a but is equal to the ratio $\frac{\theta_1 + \theta_2}{\theta_1}$, that is,

$$\frac{\mu(a, \theta_1 + \theta_2, \theta_2)}{\mu(a, \theta_1, \theta_2)} = \frac{\theta_1 + \theta_2}{\theta_1} \quad (1.30)$$

and

$$\mu(a+1, \theta_1 + \theta_2, \theta_2) = \frac{(\theta_1 + \theta_2)}{a\theta_1\theta_2} \mu(a, \theta_1, \theta_2) - \frac{(\theta_1 + \theta_2)}{\theta_2} \quad (1.31)$$

Proof

The mean $\mu(a, \theta_1, \theta_2)$ of the QNBD model is defined as:

$$\begin{aligned}
 \mu(a, \theta_1, \theta_2) &= E(X) \\
 &= \sum_{x=0}^{\infty} x \frac{(a+x-1)!}{(a-1)!x!} \frac{\theta_1(\theta_1 + x\theta_2)^{x-1}}{(1 + \theta_1 + x\theta_2)^{a+x}} \\
 &= \sum_{x=1}^{\infty} \frac{(a+x-1)!}{(a-1)!(x-1)!} \frac{\theta_1(\theta_1 + x\theta_2)^{x-1}}{(1 + \theta_1 + x\theta_2)^{a+x}}
 \end{aligned}$$

Replacing x by $(x+1)$ in the equation above results in

$$\begin{aligned}
 \mu(a, \theta_1, \theta_2) &= \\
 &= \frac{\theta_1}{(\theta_1 + \theta_2)} \sum_{x=0}^{\infty} \frac{(a+x)(\theta_1 + \theta_2 + x\theta_2)}{(1 + \theta_1 + \theta_2 + x\theta_2)} P_x(a, \theta_1, \theta_2) \\
 & \quad (1.32)
 \end{aligned}$$

Rewriting the equation as

$$\begin{aligned}
 \mu(a, \theta_1, \theta_2) &= \\
 &= \frac{\theta_1}{(\theta_1 + \theta_2)} \sum_{x=0}^{\infty} (a+x) \left(\frac{1 - (1 + \theta_1)}{1 + \theta_2 + x\theta_2} \right)^{-1} P_x(a, \theta_1, \theta_2)
 \end{aligned}$$

gives (1.30) after simplifying. Expressing (1.32) as

$$\begin{aligned} \mu(a, \theta_1, \theta_2) &= \left[\left(\frac{a\theta_1}{(\theta_1 + \theta_2)} \sum_{x=0}^{\infty} (\theta_1 + \theta_2 + x\theta_2) \right) \right. \\ &\quad \left. \times \left(\frac{(a+x)!}{a! x!} \frac{(\theta_1 + \theta_2)(\theta_1 + \theta_2 + x\theta_2)^{x-1}}{(1 + \theta_1 + \theta_2 + x\theta_2)^{a+1+x}} \right) \right] \\ &= a\theta_1 + \frac{a\theta_1\theta_2}{(\theta_1 + \theta_2)} \mu(a+1, \theta_1 + \theta_2, \theta_2) \end{aligned}$$

and rearranging the terms results in (1.31). All results shown herein for the QNBD model are also true for the NBD model by taking $\theta_2 = 0$.

Acknowledgment

This project was supported by the Research Center, College of Science, King Saud University, Riyadh, Saudi Arabia.

References

- Bilal, S. and Hassan, A. (2006). On some Models leading to quasi-negative-binomial distribution, *Journal for Korean Society for Industrial and Applied Mathematics*, 11(2), 15-29.
- Charalambides, C. A. (1990). Abel series distributions with applications to fluctuations of sample functions of stochastic processes. *Communications in Statistics - Theory and Methods*, 19, 317-335.

Consul, P. C. and Famoye, F. (2006). *Lagrangian Probability Distributions*. Birkhauser Boston, C/O Springer+Business Media Inc., 233 Spring Street, New York, NY 10013, USA.

Consul, P. C., & Jain, G. C. (1973). A generalization of the Poisson distribution. *Technometrics*, 15(4), 791-799.

Hassan, A., & Bilal, S. (2008). On some properties of quasi-negative-binomial distribution and its applications, *Journal of Modern Applied Statistical Methods*, 7(2), 616-631.

Irwin, J. O. (1968). The generalized Waring distribution applied to accident theory. *Journal of the Royal Statistical Society, Series A*, 131, 205-225.

Janardan, K. G. (1975). Markov-Polya urn-model with pre-determined strategies. *Gujarat Statistics Review*, 2(1), 17-32.

Johnson, N. L., & Kotz, S. (1992). *Univariate Discrete distributions*, 2nd Edition. New York, NY: John Wiley and Sons, Inc.

Nandi, S. B., & Das, K. K. (1994). A family of Abel series distribution. *Sankhya, B*, 56(2), 147-164.

Sen, K., & Jain, R. (1996). Generalized Markov-Polya urn-model with pre-determined strategies. *Journal of Statistical Planning and Inference*, 54, 119-133.

On Some Properties and Estimation of Size-Biased Polya-Eggenberger Distribution

Anwar Hassan
King Saud University
Riyadh, Kingdom of Saudi Arabia

Sheikh Bilal
A. S. College, Srinagar
Kashmir, India

Imtiyaz Ahmad Shah
Indian Institute of Technology-Delhi
New-Delhi, India

A size-biased version of Polya-Eggenberger distribution is introduced explicitly and by a mixture model. The proposed distribution is unimodal with positive integer moments. The recurrence relation between moments (about the origin) of the proposed distribution is established and its relationship with other distributions is discussed. Different estimation techniques are proposed to estimate the parameters of the distribution.

Key words: Polya-Eggenberger distribution, size-biased Polya-Eggenberger distribution, recurrence relation between moments, estimation techniques.

Introduction

The Polya-Eggenberger distribution (PED), introduced by Polya and Eggenberger (1923) through an urn model and further analyzed by Polya (1930), is a discrete frequency distribution that was originally considered in connection with contagious distributions. The genesis of this distribution is expressed in terms of random drawings of colored balls from an urn: Initially, it is supposed that there are a white balls and b black balls in the urn; one ball is drawn at random and then replaced together with s balls of the same color. If this procedure is repeated n times and x represents the total number of times a white ball is drawn, then the distribution of x is given by:

$$P(x = k) = \binom{n}{x} \frac{a(a+s) \dots (a+x-1)s b(b+s) \dots (b+n-x-1)s}{(a+b)(a+b+s) \dots (a+b+n-1)s} \quad (1.1)$$

Where $x = 0, 1, 2, \dots, n$ and n, a, b and s are parameters of the distribution. The distribution is known as the Polya-Eggenberger distribution with parameters (n, a, b, s) .

Taking $\alpha = (a/s)$; $\gamma = (b/s)$, results in an alternative form of (1.1) in ascending factorials as:

$$P(x = k) = \binom{n}{k} \frac{\alpha^{[x]} \gamma^{[n-x]}}{(\alpha + \gamma)^{[n]}}, \quad (1.2)$$

$$x = 0, 1, 2, \dots, n$$

Anwar Hassan is an Associate Professor in the Department of Statistics & Operations Research. Email him at: anwar.hassan2007@gmail.com or amohammadabdulkhair@ksu.edu.sa. Sheikh Bilal is a Senior Assistant Professor in the Department of Statistics. Email him at sbilal_sbilal@yahoo.com. Imtiyaz Ahmad Shah is a Research Scholar in the Department of Statistics and Operation Research. Email him at: imtiyaz3880@yahoo.co.in.

Distribution (1.2) is the most convenient form of PED for computational purposes. Another way to represent (1.1) is

$$P(X = x) = \frac{\binom{-a/s}{x} \binom{-b/s}{n-x}}{\binom{-(a+b)/s}{n}}, \quad (1.3)$$

$$x = 0, 1, 2, \dots, n,$$

and an alternative form of (1.1) in terms of parameters n ; $P = a/(a+b)$; $Q = 1 - P = b/(a+b)$ and $\delta = s/(a+b)$ is

$$P(x=x) = \binom{n}{x} \frac{\prod_{j=0}^{x-1} (P+j\delta) \prod_{j=0}^{n-x-1} (Q+j\delta)}{\prod_{j=0}^{n-1} (1+j\delta)},$$

$$x = 0, 1, 2, \dots, n. \quad (1.4)$$

It is possible for s (and therefore δ) to be negative, however s must satisfy the inequality $(a+b)+s(n-1) > 0$.

Srodka (1964) gave the recurrence relation among the moments about zero of the Polya-Eggenberger distribution (1.3) as:

$$\mu'_{r+1} = (c+b+rs)^{-1} \sum_{j=0}^r \left[\begin{matrix} an \binom{r}{j} - (a-sn) \binom{r}{j+1} \\ -s \binom{r}{j+2} \end{matrix} \right] \mu'_{r-j}.$$

$$r = 0, 1, 2, \dots$$

The r^{th} factorial moment is given by

$$\mu_{(r)} = n^{(r)} \prod_{j=0}^{r-1} \left[\frac{p+j\delta}{1+j\delta} \right] \sum_{k=r-0}^{n-r} \binom{n-r}{k-r} \frac{\left[\prod_{j=0}^{k-r-1} (P'+j\delta') \right] \times \left[\prod_{j=0}^{n-k-1} (Q'+j\delta') \right]}{\prod_{j=0}^{n-r-1} (1+j\delta')}$$

$$(1.5)$$

where

$$P' = (P+r\delta)(1+r\delta)^{-1};$$

$$Q' = Q(1+r\delta)^{-1};$$

and

$$\delta' = \delta(1+r\delta)^{-1}.$$

Specifically, the first four central moments of the Polya-Eggenberger distribution are

$$\mu'_1 = nP = \frac{na}{a+b} \quad (1.6)$$

$$\mu_2 = nPQ(1+n\delta)(1+\delta)^{-1}$$

$$\mu_3 = nPQ(Q-P)(1+n\delta)(1+2n\delta)(1+\delta)^{-1}(1+2\delta)^{-1}$$

and

$$\mu_4 = \frac{nPQ(1+n\delta) \left[(1+2n\delta)(1+3n\delta)(1-3PQ) + (n-1)\{\delta+3PQ(1+n\delta)\} \right]}{(1+\delta)(1+2\delta)(1+3\delta)}$$

Models of Size-Biased Polya-Eggenberger Distribution (SBPED)

Size-biased distributions are a special case of the more general form known as weighted distributions. First introduced by Fisher (1934) to model ascertainment bias, weighted distributions were later formalized in a unifying theory by Rao (1965). Such distributions occur naturally in practice when observations from a sample are recorded with unequal probability, such as from probability proportional to size (PPS) designs. Briefly, if a random variable X has distribution $f(x, \theta)$, with unknown parameters θ , then the corresponding weighted distribution is of the form

$$f^w(x, \theta) = \frac{w(x)f(x, \theta)}{E[w(x)]} \quad (2.1)$$

where $w(x)$ is a non-negative weight function such that $E[w(x)]$ exists.

A special case arises when the weight function is of the form $w(x) = x^\beta$. Such distributions are known as size-biased distributions of order β (Patil&Ord, 1976; Patil, 1981; Mahfoud&Patil, 1982) and are written as:

$$f^*_{\beta}(x, \theta) = \frac{x^{\beta} f(x, \theta)}{\mu'_{\beta}} \quad (2.2)$$

where $\mu'_{\beta} = \sum_x x^{\beta} f(x, \theta)$ is the β^{th} raw moment of $f(x, \theta)$.

If X is a Polya-Eggenberger variate with a probability mass function as given by (1.2), then its mean is given by (1.6). The size-biased version of X , known as the size-biased Polya-Eggenberger distribution (SBPED), can be obtained directly by taking $\beta = 1$ in (2.2) and using (1.6). The resulting equation is:

$$p(x) = \frac{x}{n\alpha} \binom{n}{x} \frac{\alpha^{[x]} \gamma^{[n-x]}}{(\alpha + \gamma)^{[n]}}$$

which can be put into the form:

$$p(x) = \binom{n-1}{x-1} \frac{(\alpha+1)^{[x-1]} \gamma^{[n-x]}}{(\alpha + \gamma + 1)^{[n-1]}}, \quad x = 1, 2, \dots, n. \quad (2.3)$$

Equation (2.3) gives the probability mass function (PMF) of the size-biased Polya-Eggenberger distribution (SBPED).

The Mixture Model

The size-biased Polya-Eggenberger distribution can also be regarded as a beta mixture of size-biased binomial distribution. The PMF of the size-biased binomial distribution is given by

$$P(x) = \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x}, \quad (2.4)$$

$$0 < p < 1, \quad x = 1, 2, \dots, n$$

and the PDF of the beta distribution of first kind is

$$f(p) = \frac{1}{\beta(\alpha, \gamma)} p^{\alpha-1} (1-p)^{\gamma-1}, \quad (2.5)$$

$$0 < p < 1.$$

Compounding (2.4) with (2.5) through the values of p results in

$$P(x) = \binom{n-1}{x-1} \frac{1}{\beta(\alpha, \gamma)} \int_0^1 p^{\alpha+x-2} (1-p)^{\gamma+n-x-1} dp$$

$$= \binom{n-1}{x-1} \frac{\beta(\alpha+x-1, \gamma+n-x)}{\beta(\alpha, \gamma)}$$

which, after simplification, gives:

$$p(x) = \binom{n-1}{x-1} \frac{\alpha^{[x]} \gamma^{[n-x]}}{(\alpha + \gamma)^{[n-1]}}, \quad (2.6)$$

$$x = 1, 2, \dots, n.$$

Replacing α with $\alpha+1$, (2.6) coincides with (2.3) and the PMF of the size-biased Polya-Eggenberger distribution can be obtained.

Structural Properties of SBPED: Unimodality

The proposed model (2.3) is unimodal according to the results of Holgate (1970).

Lemma: If the mixing distribution is non-negative, continuous and unimodal, then the resulting distribution is unimodal.

The proposed model (2.3) is unimodal because the mixing distribution is a beta distribution, which is unimodal for $\alpha > 1$ and $\gamma > 1$.

Structural Properties of SBPED: Recurrence Relation between Probabilities

Taking $x=x+1$ in (2.3) and dividing the resulting equation by (2.3), results in the ratio

$$\frac{P(x+1)}{P(x)} = \frac{(n-x)}{x} \frac{(\alpha+x)}{(\gamma+n-x-1)}$$

which gives the recurrence relation between probabilities as:

$$P(x+1) = \left[\frac{(n-x)}{x} \frac{(\alpha+x)}{(\gamma+n-x-1)} \right] P(x). \quad (3.1)$$

Structural Properties of SBPED: Recurrence Relation between Moments

Multiplying (3.1) by x^k and summing the resulting equation over x results in

$$\begin{aligned} \mu'_{k+1}(n, \alpha, \gamma) &= \sum_{x=1}^n x^k \frac{(n-1)!}{(n-x-1)!(x-1)!} \frac{(\alpha+1)^{[x]} \gamma^{[n-x]}}{(\alpha+\gamma+1)^{[n-1]}} \\ &= \frac{(n-1)(\alpha+1)}{(\alpha+\gamma+1)} \left[\sum_{x=1}^{n-1} x^k \frac{(\overline{n-1-1})!}{(\overline{n-1-x})!(x-1)!} \right. \\ &\quad \left. \times \frac{(\alpha+2)^{[x-1]} \gamma^{[\overline{n-1-x}]}}{(\alpha+\gamma+2)^{[n-1-1]}} \right] \end{aligned} \quad (3.2)$$

which reduces to

$$\mu'_{k+1}(n, \alpha, \gamma) = \frac{(n-1)(\alpha+1)}{(\alpha+\gamma+1)} \mu'_{k+1}(n-1, \alpha+1, \gamma). \quad (3.3)$$

Equation (3.3) represents the recurrence relation between moments of proposed model (2.3). In particular

$$\mu'_1 = \frac{(n-1)(\alpha+1)}{(\alpha+\gamma+1)} \quad (3.4)$$

$$\mu'_2 = \frac{(n-1)(n-2)(\alpha+1)(\alpha+2)}{(\alpha+\gamma+1)(\alpha+\gamma+2)}$$

$$\mu'_3 =$$

$$\frac{(n-1)(n-2)(n-3)(\alpha+1)(\alpha+2)(\alpha+3)}{(\alpha+\gamma+1)(\alpha+\gamma+2)(\alpha+\gamma+3)}$$

and

$$\mu'_4 =$$

$$\frac{(n-1)(n-2)(n-3)(n-4)(\alpha+1)(\alpha+2)(\alpha+3)(\alpha+4)}{(\alpha+\gamma+1)(\alpha+\gamma+2)(\alpha+\gamma+3)(\alpha+\gamma+4)}$$

Structural Properties of SBPED: Factorial Moments

Suppose $\mu'_{(k)}(n, \alpha, \gamma)$ denotes the k^{th} factorial moments of proposed model (2.3), then by definition

$$\begin{aligned} \mu'_{(k)}(n, \alpha, \gamma) &= \sum_{x=1}^n x^{(k)} \frac{(n-1)!}{(n-x)!(x-1)!} \frac{(\alpha+1)^{[x]} \gamma^{[n-x]}}{(\alpha+\gamma+1)^{[n-1]}} \\ &= \sum_{x=k}^n x \frac{(n-1)!}{(n-x)!(x-k)!} \frac{(\alpha+1)^{[x-1]} \gamma^{[n-x]}}{(\alpha+\gamma+1)^{[n-1]}} \end{aligned} \quad (3.5)$$

Taking $x = x+k-1$ in (3.5) results in

$$\begin{aligned} \mu'_{(k)} &= \sum_{x=1}^{n-k+1} (x+k-1) \\ &\quad \frac{(n-1)!}{(\overline{n-k+1-x})!(x-1)!} \frac{(\alpha+1)^{[x+k-2]} \gamma^{[\overline{n-k+1-x}]}}{(\alpha+\gamma+1)^{[n-1]}} \end{aligned}$$

and, after simplifying, gives:

$$\begin{aligned} \mu'_{(k)} &= \frac{(n-1)^{(k-1)} (\alpha+1)^{[k-1]}}{(\alpha+\gamma+1)^{[k-1]}} \\ &\quad \sum_{x=1}^{n-k+1} \left[(x+k-1) \times \frac{(\overline{n-k+1-1})!}{(\overline{n-k+1-x})!(x-1)!} \right. \\ &\quad \left. \times \frac{(\alpha+k-1+1)^{[x-1]} \gamma^{[\overline{n-k+1-x}]}}{(\alpha+k-1+\gamma+1)^{[n-k+1-1]}} \right] \end{aligned}$$

This equation, together with (3.2), gives

$$\begin{aligned} \mu'_{(k)} &= \frac{(n-1)^{(k-1)} (\alpha+1)^{[k-1]}}{(\alpha+\gamma+1)^{[k-1]}} \times \\ &\quad [\mu'_1(n-k+1, \alpha+k-1, \gamma) + (k-1)] \end{aligned}$$

Using (3.4), the k^{th} factorial moment is obtained as

$$\mu'_{(k)} = \frac{(n-1)^{(k-1)}(\alpha+1)^{[k-1]}}{(\alpha+\gamma+1)^{[k-1]}} \times \left[\frac{(n-k)(\alpha+k)}{(\alpha+\gamma+k)} + (k-1) \right] \quad (3.6)$$

and the first four factorial moments are

$$\mu'_1 = \frac{(n-1)(\alpha+1)}{(\alpha+\gamma+1)}$$

$$\mu'_{(2)} = \frac{(n-1)(n-2)(\alpha+1)(\alpha+2)}{(\alpha+\gamma+1)(\alpha+\gamma+2)} + \frac{(n-1)(\alpha+1)}{(\alpha+\gamma+1)}$$

$$\mu'_{(3)} = \frac{(n-1)(n-2)(n-3)(\alpha+1)(\alpha+2)(\alpha+3)}{(\alpha+\gamma+1)(\alpha+\gamma+2)(\alpha+\gamma+3)} + \frac{2(n-1)(n-2)(\alpha+1)(\alpha+2)}{(\alpha+\gamma+1)(\alpha+\gamma+2)}$$

and

$$\mu'_{(4)} = \frac{(n-1)(n-2)(n-3)(n-4)(\alpha+1)(\alpha+2)(\alpha+3)(\alpha+4)}{(\alpha+\gamma+1)(\alpha+\gamma+2)(\alpha+\gamma+3)(\alpha+\gamma+4)} + \frac{3(n-1)(n-2)(n-3)(\alpha+1)(\alpha+2)(\alpha+3)}{(\alpha+\gamma+1)(\alpha+\gamma+2)(\alpha+\gamma+3)}.$$

Relation with Other Distributions

Theorem 4.1

Let X be a size-biased Polya-Eggenberger variate with PMF (2.3). If $\gamma \rightarrow \infty$ such that $\alpha\gamma^{-1} = \theta$, and $n \rightarrow \infty$ such that $n\theta = \lambda$, then X tends to a size-biased Poisson distribution with parameter λ .

Theorem 4.1 Proof

The PMF of proposed model (2.3) is

$$P(x) = \left[\frac{(n-1) \dots (n-x+1)}{(x-1)!} \times \frac{(\alpha+1)(\alpha+2) \dots (\alpha+x-1)\gamma(\gamma+1) \dots (\gamma+n-x-1)}{(\alpha+\gamma+1)(\alpha+\gamma+2) \dots (\alpha+\gamma+n-1)} \right]$$

Taking the limit $\gamma \rightarrow \infty$ such that $\frac{\alpha}{\gamma} = \theta$ results in

$$P(x) = \frac{(n-1) \dots (n-x+1)}{(x-1)!} \frac{\theta^{x-1}}{(1+\theta)^{n-1}}.$$

Proceeding to limit $n \rightarrow \infty$, such that $n\theta = \lambda$, the equation reduces to a size-biased Poisson distribution with parameter λ .

Estimation

Different estimation techniques are now put forth to estimate the parameters of the proposed model. The model (2.3) has three parameters n , α and γ . The parameter n is known, whereas the remaining two parameters α and γ must be estimated.

The Moment Method

Let m'_1 , m'_2 be the sample moments (about origin) of a size-biased Polya-Eggenberger distribution (2.3). The method of moments consists in comparing the sample moments with the population moments of the distribution. The two equations thus obtained are

$$m'_1 = \frac{(n-1)(\alpha+1)}{(\alpha+\gamma+1)} \quad (5.1)$$

and

$$m'_2 = \frac{(n-1)(n-2)(\alpha+1)(\alpha+2)}{(\alpha+\gamma+1)(\alpha+\gamma+2)}. \quad (5.2)$$

Dividing (5.2) by (5.1) gives:

$$(t+1) \frac{m'_2}{m'_1} = n\alpha - 2\alpha + 2n - 4 \quad (5.3)$$

where

$$\alpha + \gamma + 1 = t. \quad (5.4)$$

From equation (5.1),

$$\alpha = \frac{m'_1 t + 1 - n}{n - 1} \quad (5.5)$$

and, from (5.3),

$$\alpha = \frac{(t+1) \frac{m'_2}{m'_1} + 4 - 2n}{n - 2}. \quad (5.6)$$

Eliminating α between (5.5) and (5.6) results in:

$$\frac{m'_1 t + 1 - n}{n - 1} = \frac{(t+1) \frac{m'_2}{m'_1} + 4 - 2n}{n - 2},$$

which, after simplification, gives the value of t as

$$t = \frac{m'_2(n-1) - n + 1}{m'_1(n-2) - n(m'_1 + m'_2) - 2m'_1 + m'_2}.$$

Substituting the value of t from this equation into (5.5), the value of α can be obtained, and after substituting the values of t and α into (5.4) the value of γ can be obtained.

Using the Mean and First Two Cell Frequencies

Taking $x=1, 2$ in the size-biased Polya-Eggenberger distribution (2.3) and equating these probabilities with their corresponding relative frequencies $\frac{f_1}{N}, \frac{f_2}{N}$ results in:

$$\frac{\gamma^{[n-1]}}{(\alpha + \gamma + 1)^{[n-1]}} = \frac{f_1}{N} \quad (5.7)$$

and

$$\frac{(\alpha + 1)\gamma^{[n-2]}}{(\alpha + \gamma + 1)^{[n-1]}} = \frac{f_2}{N},$$

where

$$N = \sum f_i. \quad (5.8)$$

Dividing (5.8) by (5.7) the ratio

$$\frac{f_2}{f_1} = \frac{(n-1)(\alpha+1)}{(\gamma+n-2)} \quad (5.9)$$

is obtained. Eliminating α between (5.9) and (5.1), the value of γ results as

$$\gamma = \frac{(n-\bar{x}-1)(n-2)(f_1+f_2)}{\bar{x}f_2 + (n-1)(\bar{x}f_1 - f_2)}.$$

Substituting the value of γ from this equation into (5.1) the value of α can be obtained.

The Method of Maximum-Likelihood

The log likelihood function of size-biased Polya-Eggenberger distribution is given by

$$\begin{aligned} \log L &= N \log \Gamma(n) - \sum f_x \log \Gamma(n-x+1) \\ &\quad - \sum f_x \log \Gamma(x+1) + \sum f_x \log \Gamma(\alpha+x) \\ &\quad - N \log \Gamma(\alpha+1) + \sum f_x \log \Gamma(\gamma+n-x) \\ &\quad - N \log \Gamma(\gamma) + N \log \Gamma(\alpha+\gamma+1) \\ &\quad - \sum f_x \log \Gamma(\alpha+\gamma+n), \end{aligned}$$

where f_x is the observed frequency for the variate value x and $N = \sum f_x$.

The proposed model has two unknown parameters, namely, (α, γ) . The log likelihood equations for estimating α and γ are

$$\frac{\partial \log L}{\partial \alpha} = 0 = -N \sum_{k=1}^{n-1} \frac{1}{\alpha + \gamma + k} + \sum f_x \sum_{k=1}^{x-1} \frac{1}{\alpha + k} \quad (5.10)$$

and

$$\frac{\partial \log L}{\partial \gamma} = 0 = \sum f_x \sum_{k=0}^{n-x-1} \frac{1}{\gamma+k} + \sum f_x \sum_{k=1}^{n-1} \frac{1}{\alpha+\gamma+k} \quad (5.11)$$

These equations do not provide direct solutions, thus an iterative solution method, such as Newton-Rampson or Fisher's scoring method, are required to solve these equations. The following system of equations may also be solved:

$$(\hat{\theta} - \theta_0) \left[\frac{\partial^2 \log L}{\partial \theta^2} \right]_{\theta_0} = \left[\frac{-\partial \log L}{\partial \theta} \right]_{\theta_0}$$

where $\hat{\theta} = (\alpha, \gamma)$ is a parameter vector, the ML estimate of θ and θ_0 is the trial value of θ which may first be obtained by equating the theoretical frequencies with the observed frequencies.

Acknowledgement

This project was supported by the Research Center, College of Science, King Saud University, Riyadh, Saudi Arabia.

References

- Bosch, A. J. (1963). The Polya distribution. *Satistica Neerlandica*, 17, 201-213.
- Eggenberger, F., & Pólya, G. (1923). Über die Statistik verketteter Vorgänge. *Zeitschrift für Angewandte Mathematik und Mechanik (ZAMM)*, 3, 279-289.

Fisher, R. A. (1934). The effects of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, 6, 13-25.

Holgate, P. (1970). The modality of compound Poisson distribution. *Biometrika*, 57, 665-667.

Mahfoud, M., & Patil, G. P. (1982). On weighted distributions. In *Statistics and probability: Essays in honour of C. R. Rao*, G. Kallianpur, P. Krishnaiah & J. Ghosh (Eds.), 479-492. New York, NY: North-Holland.

Mishra, A., & Sing, S. K. (1993). A size-biased Lagrangian Poisson distribution and its use in zero-truncated cases. *Ranchi University Mathematical Journal*, 24, 21-28.

Patil, G. K. (1981). Studies in statistical ecology involving weighted distributions. In *Applications and new directions, proceedings of Indian statistical institute golden jubilee*, J. K. Ghosh & J. Roy (Eds), 478-503. Calcutta, India: Statistical Publishing Society.

Patil, G. K., & Ord, J. K. (1976). On size-biased sampling and related form-invariant distributions. *Sankhya, Series B*, 38(1), 48-61.

Polya, G. (1930). Sur quelques points de la theorie des probabilités. *Annals de l'Institut H. Poincare*, 1, 117-161.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. In *Classical and contagious discrete distributions*, G. P. Patil (Ed.), 320-332. Calcutta, India: Statistical Publishing Society.

Srodka, T. (1963). On approximation of Hypergeometric distributions. *Zeszyty Naukowe Politechniki Lodzkiej*, 10(53), 5-17.

Modified EDF Goodness of Fit Tests for Logistic Distribution under SRS and RSS

S. A. Al-Subh
Jerash Private University,
Jerash, Jordan

M. T. Alodat
Qatar University,
Qatar

Kamaruzaman Ibrahim Abdul Aziz Jemain
Kebangsaan University,
Selangor, Malaysia

Modified forms of goodness of fit tests are presented for the logistic distribution using statistics based on the empirical distribution function (EDF). A method to improve the power of the modified EDF goodness of fit tests is introduced based on Ranked Set sampling (RSS). Data are collected via the Ranked Set Sampling (RSS) technique (McIntyre, 1952). Critical values for the logistic distribution with unknown parameters are provided and the powers of the tests are given for a number of alternative distributions. A simulation study is presented to illustrate the power of the new method.

Key words: Goodness of fit tests, empirical distribution function, power, logistic distribution, ranked set sample, Kolmogorov-Smirnov statistic.

Introduction

Many sampling methods can be used to estimate the population parameters. However, in many situations the experimental units for the variable of interest can be more easily ranked than quantified. The use of the method of ranked set sampling (RSS) in these situations is highly beneficial and is superior to simple random sampling (SRS). In many agricultural and environmental studies, it is possible to rank the experimental or sampling units with respect to the variable of interest, without actually measuring them; this usually results in cost-savings. The RSS sampling method can be used when measurements of sample units, drawn from the population of interest, are very

laborious or costly in time or money, but can be easily arranged (ranked) in order of their magnitude.

McIntyre (1952) was the first to introduce ranked set sampling (RSS). RSS gives a sample that is more informative than a simple random sample (SRS) concerning a population of interest. The RSS technique can be described as follows: Select m random samples from a population of interest each of size m . From the i^{th} sample use a visual inspection to detect the i^{th} order statistic and choose it for actual quantification, for example, Y_i , $i = 1, \dots, m$. Assuming the ranking is perfect RSS is the set of the order statistics $Y_1, \dots, Y_m$. The RSS technique can be repeated r times to obtain additional observations; these resulting measurements form an RSS of size rm .

Two factors affect the efficiency of an RSS: set size and ranking errors. The larger the set size, the larger the efficiency of RSS, while the larger the set size the more the difficulty in the visual ranking and hence the larger the ranking error (Al-Saleh & Al-Omari, 2002). Takahasi and Wakimoto (1968) provided the theoretical setups for RSS by showing that the mean of an RSS is the minimum variance unbiased estimator for a population mean. Dell and Clutter (1972) further showed that the sample mean RSS remains unbiased and more

Sameer Al-Subh is an Assistant Professor in the Department of Mathematics. Email him at: salsubh@yahoo.com. M. T. Alodat is an Associate Professor in the Department of Mathematics, Statistics and Physics, Qatar University, Qatar. Email him at: alodatmts@yahoo.com. Kamaruzaman Ibrahim is a Professor of statistics in the School of Mathematical Sciences. Email him at: kamarulz@ukm.my. Abdul Aziz Bin Jemain is a Professor of statistics in the School of Mathematical Sciences. Email him at: azizj@ukm.my.

efficient than the sample mean even if ranking is imperfect.

Several authors have modified RSS to reduce the error in ranking and to make visual ranking tractable by experimenter. (For details about RSS and its modifications, see Muttalak, 1997; Samawi, et al, 1996; Al-Odat & Al-Saleh, 2001; Bhoj, 1997; Chen, 2000; Patil, et al, 1994a).

Stockes and Sager (1988) studied the characterization of RSS. In addition, for deriving the null distribution of their proposed test, they introduced an unbiased estimator for the population distribution function based on the empirical distribution function of RSS. Also, proposed a Kolmogorov-Smirnov goodness of fit test based on the empirical distribution function (EDF). Ibrahim et al. (2011) introduced a method to improve the power of the Chi-square goodness of fit test based on RSS. They used Kullback-Leibler information to compare data collected via both SRS and RSS and conducted a simulation study for the power of Chi-square test of the new method.

Al-Subh et al. (2009) conducted a comparison study for the power of a set of EDF goodness of fit tests for the logistic distribution under SRS and RSS. This article proposes a method to improve the power of the EDF goodness of fit tests for logistic distribution under RSS and uses a simulation study to compare the powers of each test under the RSS.

MEDF Goodness of Fit Tests

Stephens (1974) presented a practical guide to goodness of fit tests using statistics based on the EDF. Green and Hegazy (1976) studied modified forms of the Kolmogorov-Smirnov D , Cramer-von Mises W^2 and Anderson-Darling A^2 goodness of fit tests. Stephens (1979) gave goodness of fit tests for the logistic distribution based on a SRS; a comprehensive survey of goodness of fit tests based on SRS can be found in Stephens (1986).

Let $X_1, X_2, \dots, X_n$ be a random sample from the distribution function $F(x)$ where $X_1 < X_2, \dots, < X_n$ is the order statistics of random sample of size n from $F(x)$. Assume

that the objective is to test the statistical hypotheses

$$H_0 : F(x) = F_o(x) \quad \forall x$$

vs.

$$H_1 : F(x) \neq F_o(x)$$

for some x , where $F_o(x)$ is a known distribution function.

The MEDF goodness of fit tests in SRS are defined as:

a) Tests related to Kolmogorov statistic, D :

$$D_1 = \max_{1 \leq i \leq n} \left| F\left(\frac{x_{(i)} - \alpha}{\beta}\right) - \left(\frac{i}{n}\right) \right|,$$

where $i = 1, 2, \dots, n$ and n is the sample size.

$$D_{11} = \max_{1 \leq i \leq n} \left| F\left(\frac{x_{(i)} - \alpha}{\beta}\right) - \left(\frac{i}{n+1}\right) \right|,$$

$$D_2 = \sum_{i=1}^n \left| F\left(\frac{x_{(i)} - \alpha}{\beta}\right) - \left(\frac{i}{n}\right) \right|,$$

$$D_{22} = \sum_{i=1}^n \left| F\left(\frac{x_{(i)} - \alpha}{\beta}\right) - \left(\frac{i+0.5}{n+1}\right) \right|,$$

$$D_3 = \sup_{1 \leq i \leq n} \left| F\left(\frac{x_{(i)} - \alpha}{\beta}\right) - \left(\frac{i}{n}\right) \right|,$$

and

$$D_4 = \sum_{i=1}^n \max \left\{ \left| \frac{i}{n} - F\left(\frac{x_{(i)} - \alpha}{\beta}\right) \right|, \left| \frac{i-1}{n} - F\left(\frac{x_{(i)} - \alpha}{\beta}\right) \right| \right\}.$$

b) Tests related to Cramer-von Mises statistic,
 W^2 :

$$W_o = \sum_{i=1}^n \left[F\left(\frac{x_{(i)} - \alpha}{\beta}\right) - \left(\frac{2i-1}{2n}\right) \right]^2,$$

$$W_{11} = \sum_{i=1}^n \left[F\left(\frac{x_{(i)} - \alpha}{\beta}\right) - \frac{i}{n+1} \right]^2,$$

$$W_{21} = \sum_{i=1}^n \left[F\left(\frac{x_{(i)} - \alpha}{\beta}\right) - \left(\frac{2i-1}{2(n+1)}\right) \right]^2.$$

c) Tests related to Anderson-Darling statistic,
 A^2 :

$$aa_{21} =$$

$$-n - \frac{n}{(r+1)^2} \left\{ \sum_{i=1}^n \left[(2i-1) \ln F\left(\frac{x_{(i)} - \alpha}{\beta}\right) + (2i+1) \ln \left(1 - F\left(\frac{x_{(n-i+1)} - \alpha}{\beta}\right)\right) \right] \right. \\ \left. - \left[(2n+1) \ln F\left(\frac{x_{(n)} - \alpha}{\beta}\right) - \ln \left(1 - F\left(\frac{x_{(n-i+1)} - \alpha}{\beta}\right)\right) \right] \right\},$$

$$aa_{22} =$$

$$-n - \left(\frac{2n}{(n+1)^2} \right) \left\{ \sum_{i=1}^n i \left[\ln F\left(\frac{x_{(i)} - \alpha}{\beta}\right) + \ln \left(1 - F\left(\frac{x_{(n-i+1)} - \alpha}{\beta}\right)\right) \right] \right\} \\ - \left(\frac{n}{(n+1)^2} \right) \left[0.25 \left\{ \ln F\left(\frac{x_{(i)} - \alpha}{\beta}\right) + \ln \left(1 - F\left(\frac{x_{(n)} - \alpha}{\beta}\right)\right) \right\} \right. \\ \left. + (n+0.75) \left\{ \ln F\left(\frac{x_{(i)} - \alpha}{\beta}\right) + \ln \left(1 - F\left(\frac{x_{(1)} - \alpha}{\beta}\right)\right) \right\} \right],$$

and

$$aa_{12} =$$

$$-(n+1) - \left(\frac{1}{n+1} \right) \left\{ \sum_{i=1}^n 2i \left[\ln F\left(\frac{x_{(i)} - \alpha}{\beta}\right) + \ln \left(1 - F\left(\frac{x_{(n-i+1)} - \alpha}{\beta}\right)\right) \right] \right\}. \quad (1)$$

This study examines the case $F_o(x) = (1 + e^{-(x-\alpha)/\beta})^{-1}$, that is, for the logistic distribution. A simulation study is conducted to show that the test T^* is more powerful than the test T when compared based on samples of the same size. The power of the T^* test can be calculated according to the equation:

$$T^*(H) = P_H(T^* > d_\alpha), \quad (2)$$

where H is a cdf under alternative hypothesis H_1^* . Here d_α is the 100α percentage point of the distribution of T^* and H_o . Due to the behavior of RSS test statistics relative to SRS

test statistics, the efficiency of the test statistics is calculated as a ratio of powers:

$$eff(T^*, T) = \frac{\text{power of } T^*}{\text{power of } T},$$

where T^* is more powerful than T if $eff(T^*, T) > 1$.

Test for Logistic Distribution

Let $X_{(1)1}, X_{(2)1}, \dots, X_{(m)r}, n = mr$ be a RSS of size $n = mr$ from a distribution function $F(x)$. The test described is an upper-tail test. A goodness of fit test is performed for the hypotheses:

$$H_0 : F(x) = F_o(x) \forall x,$$

vs.

$$H_1 : F(x) \neq F_o(x)$$

where $F_o(x) = (1 + e^{-(x-\alpha)/\beta})^{-1}$.

If α and β are unknown, then they may be estimated using their maximum likelihood estimator i.e, from $l(\alpha, \beta)$, by making the log likelihood function of the data:

$$l(\alpha, \beta) = -n \ln(\beta) - \sum_{i=1}^n (z_i) - 2 \sum_{i=1}^n \ln(1 + e^{-z_i}),$$

and in RSS by

$$\begin{aligned} l(\alpha, \beta) = & -n \ln(\beta) + \sum_{j=1}^r \sum_{i=1}^m (i-1) \ln F(z_{(i)j}) \\ & + \sum_{j=1}^r \sum_{i=1}^m (m-i) \ln(1 - F(z_{(i)j})) \\ & + \sum_{j=1}^r \sum_{i=1}^m \ln f(z_{(i)j}), \end{aligned} \quad (3)$$

where

$$z_i = (x_i - \alpha)/\beta, \quad z_{(i)j} = (x_{(i)j} - \alpha)/\beta$$

and

$$f(z_{(i)j}) = \frac{e^{-z_{(i)j}}}{\beta(1 + e^{-z_{(i)j}})^2}.$$

Using the tests given in (1) and based on the data $X_{(1)1}, X_{(2)1}, \dots, X_{(m)r}, n = mr$ called via the RSS.

Power Comparison Algorithm

Let T denote a test in (1) based on SRS and T^* be the same test, but based on RSS. To compare the power of the test T^* with the power of the test T based on samples of the same size, first the algorithm to calculate the percentage points is introduced:

1. Let $x_{(i)j}$ be a random sample from $F_o(x)$.
2. Estimate parameters α and β from the sample by maximum likelihood; the estimates are given by (3).
3. Find the EDF $F_n^*(x)$ as follows:

$$\begin{aligned} F_n^*(x) &= \frac{1}{mr} \sum_{j=1}^r \sum_{i=1}^m I(x_{(i)j} \leq x), \\ I(x_{(i)j} \leq x) &= \begin{cases} 1, & x_{(i)j} \leq x, \\ 0, & \text{o.w.} \end{cases} \end{aligned} \quad (4)$$

4. Use $F_n^*(x)$ to calculate the value of T^* as in (1).
5. Repeat steps one through four 10,000 times to obtain $T_1^*, \dots, T_{10,000}^*$.
6. The percentage point d_α of T^* is approximated by the $(1-\alpha)100$ quantile of $T_1^*, \dots, T_{10,000}^*$.

The following algorithm is designed to obtain the power of T^* at a distribution, for example, H , under H_o :

1. Let $x_{(i)j}$ be a random sample from $F_o(x)$.
2. Estimate the parameters α, β from the sample by maximum likelihood; the estimates are given by (3).
3. Find the EDF $F_n^*(x)$ as in (4).
4. Calculate the value of T^* in (1).
5. Repeat steps one through four 10,000 times to obtain $T_1^*, \dots, T_{10,000}^*$.
6. Calculate the power of

$$T^*(H) \approx \frac{1}{10,000} \sum_{t=1}^{10,000} I(T_t^* > d_\alpha),$$

where $I(\cdot)$ stands for indicator function.

Results

A simulation study was conducted to compare the power of T and T^* . The power, as well as the percentage point, of each test are approximated based on a Monte Carlo simulation of 10,000 iterations according to the algorithm described previously. Table 1 shows the percentage points for the 5% level for the null hypotheses of the logistic distribution for RSS. The efficiency of the tests was compared for different sized samples: $r = 3, 5, 10, 25$; different set sizes: $m = 2, 3$; and different alternative distributions: $Normal = N(\alpha, \beta^2)$, $Laplace = L(\alpha, \beta)$, $Cauchy = C(\alpha, \beta)$, $StudentT = S(5)$, $Uniform = U(\alpha, \beta)$, and $Lognormal = LN(\alpha, \beta)$. Comparisons were made only for cases where the data are quantified via RSS. Simulation results are shown in the Tables 2 and 3. For the lognormal and uniform distributions, computations show

that the powers of all test statistics equal one, thus, these powers are not reported.

Based on study results, the following conclusions are put forth:

1. The efficiencies in Tables 1 and 3 are all greater than 1; this indicates that the MEDF tests under ERSS are more powerful than their counterparts in SRS.
2. Tables 1-3 show that the efficiency increases as the distribution under the alternative hypothesis departs to asymmetry.
3. Power increases as the sample size n increases.
4. Power is equal to one for the lognormal and uniform distributions.
5. The MEDF tests based on data collected via RSS are more powerful than the EDF tests based on an SRS of the same size.

Conclusion

The power of a set of modified EDF goodness of fit tests was shown to be improved if a sample is collected via the RSS method, as opposed to the SRS method. Moreover, modified EDF tests show excellent power performance in comparison to their SRS counterparts. Although this study is limited to the logistic distribution under the null hypothesis, it could be easily extended to other distributions.

References

- Al-Odat, M. T., & Al-Saleh, M. F. (2001). A variation of ranked set sampling. *Journal of Applied Statistical Science*, 10, 137-146.
- Al-Saleh, M. F., & Al-Omari, A. I. (2002). Multistage ranked set sampling. *Journal of Statistical planning and Inference*, 102, 273-286.
- Ibrahim, K., Alodat, M. T., Jemain, A. A., & Al-Subh, S. A. (2011). Chi-square test for goodness of fit for logistic distribution using ranked set sampling and simple random sampling. *Statistica & Applicazioni*, IX(2), 111-128.

Table 1: Percentage Points for SRS and RSS, $\alpha = 0.05$

Test	SRS					RSS				
	n					n				
	6	10	20	30	50	6	10	20	30	50
D_1	0.269	0.215	0.158	0.132	0.103	0.271	0.221	0.166	0.140	0.116
D_{11}	0.198	0.175	0.141	0.122	0.099	0.200	0.175	0.143	0.126	0.107
D_2	0.723	0.885	1.191	1.460	1.861	0.747	0.932	1.312	1.631	2.249
D_{22}	0.693	0.881	1.201	1.470	1.867	0.712	0.884	1.244	1.537	2.128
D_3	0.287	0.228	0.168	0.140	0.110	0.291	0.237	0.177	0.150	0.122
D_4	1.096	1.283	1.618	1.903	2.311	1.115	1.322	1.731	2.062	2.694
WW_0	0.081	0.087	0.091	0.097	0.096	0.087	0.098	0.109	0.118	0.137
WW_{11}	0.085	0.089	0.093	0.098	0.096	0.088	0.089	0.098	0.104	0.122
WW_{21}	0.116	0.110	0.104	0.105	0.101	0.118	.110	0.109	0.112	0.127
aa_{21}	0.261	0.376	0.488	0.550	0.593	0.312	0.393	0.499	0.549	0.690
aa_{22}	0.745	0.757	0.726	0.727	0.711	0.761	0.754	0.740	0.751	0.850
aa_{12}	0.421	0.513	0.577	0.618	0.641	0.460	0.505	0.573	0.623	0.762

Table 2: Efficiency Values of Tests Using RSS with respect to SRS for $n = 6, 10, 20, 30, 50$ and $\alpha = 0.05$

H	T	n				
		6	10	20	30	50
$N(\theta, \sigma^2)$	D_1	5.52	1.55	1.16	1.01	1
	D_{11}	0.27	1.98	1.49	1.06	1
	D_2	8.67	1.21	0.98	1	1
	D_{22}	0.36	2.05	1.11	1	1
	D_3	1.42	1.39	1.08	0.99	1
	D_4	1.97	1.18	0.99	1	1
	WW_0	1.95	1.34	0.99	1	1
	WW_{11}	1.54	2.29	0.94	1	1
	WW_{21}	1.25	2.17	0.94	1	1
	aa_{21}	2.12	2.17	1	1	1
	aa_{22}	4.38	4.91	0.98	1	1
	aa_{12}	1	3.21	1.52	1	1
$L(\theta, \sigma)$	D_1	1.38	1.43	1.4	1.21	1.04
	D_{11}	1.19	1.38	1.42	1.24	1.05
	D_2	1.08	1.43	1.27	1.04	1
	D_{22}	0.85	1.35	1.7	1.31	1
	D_3	1.05	1.16	1.18	1.11	1.02
	D_4	1.51	1.65	1.26	1.02	1
	WW_0	1.26	1.4	1.17	1.03	1
	WW_{11}	0.83	1.51	1.6	1.21	1
	WW_{21}	0.51	0.94	1.47	1.21	1
	aa_{21}	1.15	1.66	2.11	1.52	1
	aa_{22}	0.82	1.48	2.17	1.53	1
	aa_{12}	0.47	0.95	2.19	1.57	1

Table 2 (continued): Efficiency Values of Tests Using RSS with respect to SRS for $n = 6, 10, 20, 30, 50$ and $\alpha = 0.05$

H	T	n				
		6	10	20	30	50
$C(\theta, \sigma)$	D_1	1.7	2.07	2.17	2.09	1.75
	D_{11}	0.72	0.91	1.14	1.26	1.27
	D_2	1.28	1.41	1.56	1.68	1.56
	D_{22}	1.19	1.41	1.52	1.62	1.5
	D_3	1.03	1.26	1.42	1.54	1.42
	D_4	0.68	0.9	1.09	1.29	1.32
	WW_0	0.76	0.98	1.16	1.37	1.37
	WW_{11}	0.71	0.93	1.14	1.32	1.3
	WW_{21}	0.49	0.47	0.66	0.91	1.11
	aa_{21}	0.85	0.91	1	1.1	1.25
	aa_{22}	0.89	0.96	1.05	1.13	1.25
	aa_{12}	0.86	0.95	1.07	1.12	1.26
$S(5)$	D_1	1.97	1.75	1.59	1.25	1.01
	D_{11}	2.91	2.48	2.12	1.43	1.03
	D_2	1.96	2.13	1.11	0.98	1
	D_{22}	1.15	2.7	2.57	1.1	1
	D_3	1.8	1.48	1.32	1.13	1
	D_4	2.53	2.35	1.06	0.99	1
	WW_0	2.2	2	1.12	0.98	1
	WW_{11}	1.64	3.46	2.35	1.1	1
	WW_{21}	0.74	2.2	2.26	1.11	1
	aa_{21}	2.34	3.37	2.7	1.02	1
	aa_{22}	1.16	3.02	3.32	1.07	1
	aa_{12}	0.75	1.45	4.03	1.48	1

Table 3: 1,000×Power Values for SRS and RSS-Two Unknown Parameters, $\alpha = 0.05$

H	Test	SRS					RSS				
		n					n				
		6	10	20	30	50	6	10	20	30	50
$N(\theta, \sigma^2)$	D_1	135	306	750	976	1000	254	473	872	983	1000
	D_{11}	42	115	522	911	1000	52	228	780	970	1000
	D_2	184	492	998	1000	1000	260	595	1000	1000	1000
	D_{22}	29	132	822	1000	1000	47	271	916	1000	1000
	D_3	125	336	819	1000	1000	245	466	887	1000	1000
	D_4	172	579	1000	1000	1000	339	685	1000	1000	1000
	WW_0	165	494	1000	1000	1000	330	661	1000	1000	1000
	WW_{11}	28	129	807	1000	1000	43	296	941	1000	1000
	WW_{21}	33	134	810	1000	1000	45	291	940	1000	1000
	aa_{21}	108	253	1000	1000	1000	235	550	1000	1000	1000
	aa_{22}	8	56	1000	1000	1000	35	275	1000	1000	1000
	aa_{12}	1	19	560	1000	1000	1	61	907	1000	1000
$L(\theta, \sigma)$	D_1	192	295	519	721	938	281	434	729	876	980
	D_{11}	98	211	473	687	930	117	296	673	852	977
	D_2	280	317	642	918	1000	308	465	818	956	1000
	D_{22}	205	218	414	705	1000	182	295	703	921	1000
	D_3	281	394	632	800	963	314	469	744	888	982
	D_4	164	272	660	941	1000	252	448	831	964	1000
	WW_0	200	335	715	930	1000	269	475	838	962	1000
	WW_{11}	89	166	459	769	998	78	256	734	931	1000
	WW_{21}	165	214	447	747	997	90	201	658	907	1000
	aa_{21}	170	199	340	621	1000	217	331	718	941	1000
	aa_{22}	115	146	304	603	1000	101	216	660	921	1000
	aa_{12}	74	108	268	573	997	35	103	586	899	1000

EDF GOODNESS OF FIT TESTS FOR LOGISTIC DISTRIBUTION UNDER SRS AND RSS

Table 3 (continued): 1,000×Power Values for SRS and RSS-Two Unknown Parameters, $\alpha = 0.05$

H	Test	SRS					RSS				
		n					n				
		6	10	20	30	50	6	10	20	30	50
$C(\theta, \sigma)$	D_1	167	150	211	264	397	284	310	463	552	693
	D_{11}	348	340	360	420	527	255	321	416	530	669
	D_2	348	321	338	359	472	444	453	533	602	738
	D_{22}	365	341	380	403	521	436	480	583	663	782
	D_3	271	223	287	320	448	280	281	407	492	634
	D_4	326	297	349	367	490	221	268	380	474	646
	WW_0	314	281	339	367	492	238	274	393	503	672
	WW_{11}	351	324	391	423	554	248	300	450	558	719
	WW_{21}	321	326	379	415	536	157	156	255	390	595
	aa_{21}	628	685	756	742	656	534	620	753	808	817
	aa_{22}	683	743	789	757	660	608	710	825	852	826
	aa_{12}	700	748	777	761	658	600	707	829	856	827
$S(5)$	D_1	98	175	430	702	980	213	356	690	879	990
	D_{11}	11	56	268	583	957	32	168	569	831	987
	D_2	95	163	756	999	1000	196	402	840	1000	1000
	D_{22}	39	57	267	861	1000	59	173	685	950	1000
	D_3	113	206	520	777	991	208	360	685	879	991
	D_4	71	164	820	1000	1000	205	441	868	1000	1000
	WW_0	81	186	770	1000	1000	207	435	865	1000	1000
	WW_{11}	11	44	304	869	1000	23	167	715	958	1000
	WW_{21}	32	64	307	860	1000	25	151	694	953	1000
	aa_{21}	59	83	294	960	1000	160	328	805	980	1000
	aa_{22}	32	50	208	899	1000	37	159	706	966	1000
	aa_{12}	20	31	132	638	1000	15	50	585	947	1000

- Al-Subh, S. A., Alodat, M. T., Ibrahim, K., & Jemain, A. A. (2009). EDF goodness of fit tests of logistic distribution under selective order statistics. *Pakistan Journal of Statistics*, 25(3), 265-274.
- Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1992). *A first course in order statistics*. New York, NY: John Wiley and Sons.
- Balakrishnan, N. (1992). *Handbook of the logistic distribution*. New York, NY: Marcel Dekker, Inc.
- Bhoj, D. S. (1997). Estimation of parameters of the extreme value distribution using ranked set sampling. *Communications in Statistics-Theory and Methods*, 26(3), 653-667.
- Chen, Z. (2000). On ranked-set sample quantiles and their applications. *Journal of Statistical Planning and Inference*, 83, 125-135.
- D'Agostine, R., & Stephens. M. (1986). *Goodness of fit techniques*. New York, NY: Marcel Dekker, Inc.
- Dell, D. R., & Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28, 545-555.
- Green J. R., & Hegazy, Y. A. S. (1976). Powerful modified-EDF goodness-of-fit tests. *Journal of the American Statistical Association*, 71(353), 204-209.
- Ibrahim, K. Alodat, M. T. Jemain, A. A., & Al-Subh, S. A (2011). Chi-square test for goodness of fit for logistic distribution using ranked set sampling and simple random sampling. *Statistica & Applicazioni*, IX(2), 111-128.
- McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, 3, 385-390.
- Muttlak, H. A. (1997). Median ranked set sampling. *Journal of Applied Statistical Science*, 6, 245-255.
- Patil, G. P., Sinha, A. K., & Taillie, C. (1994a). Ranked set sampling. In *Handbook of statistics, 12: Environmental statistics*, G.P. Patil & C. R. Rao (Eds.), 167-200. New York: North Holland Elsevier.
- Samawi, H. M., Ahmed, M. S., & Abu-Dayyeh, W. (1996). Estimation the population mean using extreme ranked set sampling. *Biometrical Journal*, 38, 577-586.
- Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69, 730-737.
- Stephens, M. A. (1979). Tests of fit for the logistic distribution based on the empirical distribution function. *Biometrika*, 66(3), 591-595.
- Stockes S. L., & Sager T. W. (1988). Characterization of a ranked -set sample with application to estimating distribution functions. *Journal of the American Statistical Association*, 83(402), 374-381.
- Takahasi, K., & Wakitmoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 20, 1-31.

Single Sampling Plans for Variables Indexed by AQL and AOQL with Measurement Error

R. Sankle J. R. Singh
Vikram University,
India, Ujjain (M. P.)

Single sampling plans are investigated for variables indexed by acceptable quality level (AQL) and average outgoing quality limit (AOQL) under measurement error. Procedures and tables are provided for selection of single sampling plans for variables for given AQL and AOQL when rejected lots are 100% inspected for replacement of a nonconforming unit. For a particular sampling plan in operation for an observed measurement, a method for determining true operating characteristic (OC) functions and average outgoing quality (AOQ) is described for various error sizes.

Key words: Measurement error, AQL, AOQL.

Introduction

One difficulty with production processes is achieving a desired quality level of manufactured product while maintaining economy in production cost. Statistical techniques have been successfully applied to address this problem; to employ statistical techniques, inspections are conducted on intermediate and finished products. In every inspection system, there is always a possibility for error in accepting a non-conforming unit and rejecting a conforming unit. These errors, which are mainly due to chance, are termed inspection errors and they can be estimated. This is important because corrective action must be taken if the number of inspection errors is large. Jackson (1957) studied the effect of inspection errors on waste and on the quality of outgoing product assuming 100% inspection. Considering that error is a substantial part of observed variation, Diviney and David (1963)

investigated the relationship between measurement error and product acceptance.

The requirement that the measurement of an individual item does not exceed some specified limit is sometimes more important than the requirement that the mean and variability for the items be at or near some pre-determined value. An acceptance sampling plan in which a specified number of units is sampled from each lot, with the lot being accepted if less than a fixed number of non-conformance products are found in the sample, is one of the traditional statistical tools used for quality control. Lots that are not accepted can either be discarded or rectified. Rectification, that is, replacing or discarding all non-conforming units after 100% inspection of rejected lots, is frequently used when manufacturing costs are high.

Several authors have proposed predictors for estimating the number or rate of non-conformances in lots subjected to acceptance sampling (Hahn, 1986; Zaslavsky, 1988; Brush, et al. 1990; Martz & Zimmer, 1990). Greeberg and Stokes (1992) used the information obtained in rectification to devise a more efficient predictor than those previously proposed. Greenberg and Stockes (1995) also considered an application of quality control in which the test procedure is imperfect. Two problems may exist in acceptance sampling. Devices that are classified as non-conforming may be conforming (false positive) and devices

R. Sankle is on the faculty in the School of Studies in Statistics. Email her at: rajshreesankle@gmail.com. J. R. Singh is a Professor in the School of Studies in Statistics. Email him at: jrsinghstat@gmail.com.

that are classified as conforming may be non-conforming (false negative). Johnson, et al. (1991) provided expressions and tables of the average outgoing quality for many types of sampling plans when the false positive and false negative rates are known. Lindsay (1985) described methods for estimating the probability of false positives and false negatives and the rates and numbers of non-conformances when a sample is repeatedly inspected. However, these authors do not consider plans with rectification.

A lot-by-lot rectification inspection scheme for a series of lots calls for 100% inspection of rejected lots under the application of a sampling plan. If it is preferable to use a single sampling plan for variables under a rectification inspection scheme, the index for the selection of the sampling plan will be the average outgoing quality limit (AOQL), which is the worst average quality the consumer will receive in the long run, regardless of the incoming quality. Rejected lots are often a nuisance to the producers because they result in extra work and extra cost. If too many lots are rejected the reputation of the producer or supplier may be damaged. From the producer's point of view, it is preferable to fix an acceptable quality level (AQL) by designing a sampling plan such that, if the incoming product quality is maintained at AQL most of the lots, for example 95%, will be accepted during the sampling inspection stage. Thus, designing sampling inspection plans indexed by AQL and AOQL satisfies both the producer and consumer whenever rectifying inspection is necessary. The predictors are generally assumed to be measured without error, but this is often not the case.

To identify the parameter in the model, the following assumptions are made concerning measurement errors. First, it is assumed that the true values and the measurement errors are uncorrelated and that the mean of the measurement errors is zero. Second, the measurement errors are assumed to be normally distributed with zero mean and a constant known variance. Third, the true values are assumed to be normally distributed with a mean estimated by the mean of the observed values and a variance estimated using the reliability of the observed values. The reliability of a variable measured with error is the ratio of the variance

of the true values to the variance of the observed values; the closer this ratio is to 1 the more reliable the measurement. The reliability can be provided by reliability coefficients (Hand, 2004). Alternatively, a range of plausible reliabilities can be explored to carry out a sensitivity analysis of the results to estimate the severity of the unobserved measurement error.

This study examines single sampling plans for variables indexed by AQL and AOQL under measurement error. Procedures and tables are provided for selecting single sampling plans for variables for given AQL and AOQL when rejected lots are 100% inspected for replacement of nonconforming units. For a particular sampling plan in operation for observed measurement, a method of determining the OC function and AOQ curves is described for various errors sizes.

Model description for Variable Single Sampling Plan indexed by AQL and AOQL under Measurement Errors:

Consider the distribution of the true quality characteristics x to be normal with mean μ and known standard deviation σ_p . The density function is:

$$f(x) = \frac{1}{\sigma_p} \Phi \left(\frac{x - \mu}{\sigma_p} \right), \quad (2.1)$$

where $\Phi(x)$ is the standardized normal probability density function given by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \quad (2.2)$$

The mean and standard deviation of the observed measurement ($X = x + e$) can be written as

$$E(X) = E(x) + E(e) = \mu$$

where μ is the mean of x and e is the random error at measurement and is independent of x , and

$$V(X) = V(x) + V(e) = \sigma_p^2 + \sigma_e^2 = \sigma_X^2.$$

SAMPLING PLANS FOR AQL AND AOQL VARIABLES WITH MEASUREMENT ERROR

The correlation coefficient ρ between the true and observed measurement is given by

$$\rho = \frac{E\{(x-\mu)(X-\mu)\}}{\sigma_p \sigma_X} = \frac{E\{(x-\mu)^2 + (x-\mu)e\}}{\sigma_p \sigma_X}.$$

Noting that x and e are independent, $E(e) = 0$ and $E(x) = \mu$, it can be shown that

$$\begin{aligned} \rho &= \frac{\sigma_p^2}{\sigma_p \sigma_X} \\ &= \frac{\sigma_p}{\sigma_X}. \end{aligned} \quad (2.3)$$

The relation between the size of measurement error r and correlation coefficient ρ is:

$$\rho = \frac{r}{\sqrt{1+r^2}} \quad (2.4)$$

where

$$r = \frac{\sigma_p}{\sigma_e}.$$

In referencing a single sampling variable plan when σ_p is known, the following symbols are used:

L: Lower specification limit;

U: Upper specification limit;

N: Sample size;

k : Acceptance parameter; and

$\bar{x}$: Sample mean

$$\Phi(y) = \int_{-\infty}^y \left(\frac{1}{\sqrt{2\pi}} \right) \exp\left(-\frac{1}{2}z^2\right) dz, \quad (2.5)$$

where $z \sim N(0,1)$.

The acceptance criterion for the single sampling plan is: For the upper specification limit, accept the lot if,

$$\bar{x} + k\sigma_p \leq U, \quad (2.6)$$

and, for the lower specification limit, accept the lot if

$$\bar{x} - k\sigma_p \geq L. \quad (2.7)$$

The fraction nonconforming in a given lot is

$$\Phi(-K_p) = p, \quad (2.8)$$

with

$$K_p = \frac{U - \mu}{\sigma_p} \quad (2.9)$$

where K_p is the p percent point of the standard normal distribution. If p is the proportion defective in the lot, then

$$U = \mu + K_p \sigma_p \quad (2.10)$$

and its probability of acceptance under measurement error will be

$$P_a(p) = \Phi(w), \quad (2.11)$$

with

$$w = (K_p - k) \frac{\sqrt{n}}{\rho}. \quad (2.12)$$

If the quality of the accepted lot is p and all non-conforming units found in the rejected lots are replaced by conforming units in a rectification inspection scheme, the AOQ can be approximated as

$$AOQ = p.P_a(p). \quad (2.13)$$

If p_m is the proportion non-conforming at which AOQ is maximum, then

$$AOQL = p_m.P_a(p_m). \quad (2.14)$$

If AQL (p_l) is prescribed, then the corresponding value of K_{AQL} or K_1 will be fixed, and if $P_a(p_l)$ is fixed at 95%, then, $w_{AQL} = w_1 = 1.645$; hence,

$$1.645 = (K_1 - k) \frac{\sqrt{n}}{\rho}, \quad (2.15)$$

so that for a given AQL, k is determined by the sample size n .

Results

Table 1 is used for selecting a single sampling variables plan under measurement errors for known σ case. For example, if the AQL is fixed at 1%, the AOQL is fixed at 1.25% and $r = 2, 4, 6$ and ∞ , Table 1 yields $n = 39, 27, 26$ and 25 , and $k = 1.989, 1.990, 1.992$ and 1.998 , respectively. It shows that, when the size of the error increases, the value of n increases and, due to measurement errors, the sample sizes are affected but there is a very minor change in acceptance parameter k . Further, suppose that it is decided to use σ , an acceptance criterion where σ_p is known to be 2.0. Let there exist an upper specific limit $U = 10.0$ and a unit for which the quality characteristic $x > U$ is considered as nonconforming.

Table 2 shows the performance characteristics of a sampling plan with $n = 25$ and $k = 2.0$ under a rectifying inspection scheme. If the true process average quality is operating at AQL ($\mu = 5.346$) and $r = \infty$, then 95% of the lots submitted will be accepted during the sampling inspection stage itself and only 5% of the rejected lots will be rectified by replacing non-conforming units with conforming units. In such a case, the AOQ will be only about 1%. If the submitted quality deteriorates to 1.79 % (error free case, that is, $r = \infty$), then only about 70% of the lots will be accepted by the sampling plan and approximately one out of every three lots will be rejected and rectified. The AOQ in such a case will not exceed the AOQL of 1.25% fixed, meaning that, irrespective of the product quality submitted by the producer, the consumer will receive an average quality not worse than 1.25% under the

rectification scheme. The worst case is when $r = 2$; the AOQ in such a case will just exceed the AOQL of 1.25% fixed for different errors sizes. The values of $P_a(p_m)$ for known σ case are presented in Table 3, and a visual comparison of AOQ curves for different error sizes is shown in Figure 1. As Figure 1 illustrates, the effect of measurement error is serious on the AOQ curves.

When using Table 1 to select sampling plans, limitations of plans indexed by AOQL under measurement error must be taken into account. Sampling with rectification of rejected lots reduces the average percentage of nonconforming items in the lots; however, it also introduces non-homogeneity in the series of lots finally accepted. That is, any particular lot will have a quality of $p\%$ or 0% non-conforming depending on whether the lot is accepted or rectified. Thus, the assumption underlying the AOQL principle is that the homogeneity in the qualities of individual lots is unimportant and only average quality matters.

Table 3 gives $P_a(p_m)$ values for the plans given in Table 1. If AQL is 0.25%, AOQL is 1.25% and $r = 2, 4, 6$ and ∞ , then $P_a(p_m)$ is 0.354, 0.342, 0.340 and 0.338, respectively, and $p_m = AOQL/P_a(p_m)$ for $r = 2, 4, 6$ and ∞ , is 3.53%, 3.65%, 3.67%, 3.69%. Thus, if the lot quality is 3.69% then, on average, among every three lots passed on to the consumer two will be free from non-conforming items while the third lot will contain 3.69% non-conforming items: this is about 15 times the AQL specified. In order to avoid such error, the producer should maintain the process quality approximately at the set AQL because a high rate of rejecting lots at $p = p_m$ will also indirectly put pressure on the producer to improve the submitted quality.

References

- Brush, G. G., Hoadley, B., & Saperstein B. (1990). Estimating outgoing quality using the quality measurement plan. *Technometrics*, 32, 31-41.
- Diviney, T. E., & David, N. A. (1963). A note on the relationship between measurement error and product acceptance. *Journal of Industrial Engineering*, 14, 218-219.

SAMPLING PLANS FOR AQL AND AOQL VARIABLES WITH MEASUREMENT ERROR

Table 1: Single Sampling Plans for Variables Indexed by AQL and AOQL under Measurement Error

r	AOQL (%)	AQL (%)											
		0.040	0.065	0.100	0.150	0.250	0.400	0.650	1.000	1.500	2.500	4.000	6.500
2	0.040	74, 3.107											
	0.080	19, 2.905	75, 2.971										
	0.125	10, 2.753	19, 2.765	63, 2.826									
	0.200	7, 2.604	10, 2.595	17, 2.606	43, 2.654								
	0.320	5, 2.458	6, 2.438	9, 2.429	14, 2.437	47, 2.504							
	0.500	3, 2.317	4, 2.289	5, 2.271	8, 2.263	14, 2.276	49, 2.350						
	0.800	3, 2.160	3, 2.235	4, 2.104	5, 2.083	7, 2.077	12, 2.091	47, 2.174					
	1.250		2, 1.965	3, 1.937	3, 1.914	4, 1.894	6, 1.888	12, 1.907	39, 1.989				
	2.000			2, 1.748	2, 1.722	3, 1.694	4, 1.677	6, 1.672	10, 1.692	25, 1.760			
	3.200					2, 1.477	3, 1.454	2, 1.438	5, 1.436	7, 1.453	26, 1.549		
	5.000							2, 1.143	3, 1.186	4, 1.186	7, 1.214	24, 1.326	
	8.000								2, 0.889	2, 0.882	3, 0.887	6, 0.923	21, 1.058
4	0.040	52, 3.109											
	0.080	17, 2.936	51, 2.969										
	0.125	10, 2.800	17, 2.795	45, 2.828									
	0.200	6, 2.665	9, 2.643	15, 2.639	33, 2.662								
	0.320	4, 2.532	6, 2.500	8, 2.481	12, 2.475	34, 2.508							
	0.500	3, 2.402	4, 2.365	5, 2.337	7, 2.318	12, 2.312	34, 2.351						
	0.800	2, 2.258	3, 2.217	4, 2.184	4, 2.157	7, 2.133	11, 2.130	32, 2.175					
	1.250		2, 2.067	3, 2.031	3, 2.000	4, 1.967	6, 1.946	10, 1.946	27, 1.990				
	2.000			2, 1.857	2, 1.823	3, 1.784	4, 1.754	5, 1.734	9, 1.734	19, 1.770			
	3.200					2, 1.585	2, 1.550	3, 1.521	4, 1.504	7, 1.503	19, 1.555		
	5.000							2, 1.299	3, 1.275	4, 1.260	6, 1.264	17, 1.327	
	8.000								2, 1.003	2, 0.982	3, 0.968	5, 0.977	15, 1.055

Table 1 (continued): Single Sampling Plans for Variables Indexed by AQL and AOQL under Measurement Error

r	AOQL (%)	AQL (%)											
		0.040	0.065	0.100	0.150	0.250	0.400	0.650	1.000	1.500	2.500	4.000	6.500
6	0.040	49, 3.111											
	0.080	17, 2.942	48, 2.971										
	0.125	10, 2.809	16, 2.802	42, 2.829									
	0.200	6, 2.677	9, 2.653	14, 2.646	32, 2.666								
	0.320	4, 2.546	6, 2.513	8, 2.491	12, 2.483	33, 2.511							
	0.500	3, 2.419	4, 2.380	5, 2.350	7, 2.329	12, 2.320	32, 2.354						
	0.800	2, 2.277	3, 2.235	4, 2.199	4, 2.171	6, 2.145	11, 2.138	30, 2.174					
	1.250		2, 2.088	3, 2.050	3, 2.017	4, 1.981	6, 1.958	10, 1.954	26, 1.992				
	2.000			2, 1.878	2, 1.843	3, 1.802	4, 1.770	5, 1.747	8, 1.744	18, 1.773			
	3.200					2, 1.606	2, 1.569	3, 1.537	4, 1.518	7, 1.514	18, 1.558		
	5.000							2, 1.320	3, 1.293	4, 1.276	6, 1.275	16, 1.329	
	8.000								2, 1.026	2, 1.002	3, 0.984	5, 0.989	14, 1.057
∞	0.040	47, 3.084											
	0.080	17, 2.900	46, 2.972										
	0.125	9, 2.754	16, 2.807	40, 2.831									
	0.200	6, 2.608	9, 2.661	14, 2.652	30, 2.670								
	0.320	4, 2.464	6, 2.523	8, 2.500	12, 2.489	31, 2.513							
	0.500	3, 2.324	4, 2.392	5, 2.361	7, 2.338	12, 2.327	31, 2.355						
	0.800	2, 2.168	3, 2.249	4, 2.213	4, 2.183	6, 2.154	11, 2.145	29, 2.178					
	1.250		2, 2.104	3, 2.065	3, 2.031	4, 1.994	6, 1.968	10, 1.961	25, 1.994				
	2.000			2, 2.006	2, 1.860	3, 1.817	4, 1.783	5, 1.748	8, 1.741	18, 1.788			
	3.200					2, 1.623	2, 1.585	3, 1.551	4, 1.529	6, 1.522	17, 1.561		
	5.000							2, 1.337	3, 1.308	3, 1.288	6, 1.284	15, 1.332	
	8.000								2, 1.044	2, 1.019	3, 1.998	5, 0.999	13, 1.060

SAMPLING PLANS FOR AQL AND AOQL VARIABLES WITH MEASUREMENT ERROR

Table 2: Performance Characteristics of the Variables Plan under Measurement Error
for AQL = 0.01, AOQL = 0.0125, U = 10, SD = 2

r	μ	v'	p(%)	w	Pa	AOQ
2	5.4330	2.2835	1.12	1.6450	0.9500	1.0640
	5.6000	2.2000	1.39	1.1785	0.8807	1.2245
	5.8000	2.1000	1.79	0.6200	0.7324	1.3083
	5.9000	2.0500	2.02	0.3407	0.6333	1.2782
	6.0000	2.0000	2.28	0.0614	0.5245	1.1932
	6.2000	1.9000	2.87	-0.4971	0.3096	0.8889
4	5.3673	2.3163	1.03	1.6450	0.9500	0.9757
	5.4000	2.3000	1.07	1.5626	0.9409	1.0091
	5.6000	2.2000	1.39	1.0586	0.8551	1.1889
	5.8000	2.1000	1.79	0.5545	0.7104	1.2690
	6.0000	2.0000	2.28	0.0504	0.5201	1.1832
	6.2000	1.9000	2.87	-0.4537	0.3250	0.9334
6	5.3618	2.3191	1.02	1.6450	0.9500	0.9686
	5.4000	2.3000	1.07	1.5491	0.9393	1.0073
	5.6000	2.2000	1.39	1.0462	0.8523	1.1849
	5.8000	2.1000	1.79	0.5432	0.7065	1.2621
	6.0000	2.0000	2.28	0.0402	0.5160	1.1740
	6.2000	1.9000	2.87	-0.4627	0.3218	0.9240
∞	5.3460	2.3270	1.00	1.6450	0.9500	0.9484
	5.4000	2.3000	1.07	1.5100	0.9345	1.0021
	5.8000	2.1000	1.79	0.5100	0.6950	1.2415
	6.0000	2.0000	2.28	0.0100	0.5040	1.1466
	6.2000	1.9000	2.87	-0.4900	0.3121	0.8961
	6.4000	1.8000	3.59	-0.9900	0.1611	0.5788

Table 3: Pa(pm) Values of Known Sigma Plans Under Measurement Error

r	AOQL (%)	AQL (%)											
		0.040	0.065	0.100	0.150	0.250	0.400	0.650	1.000	1.500	2.500	4.000	6.500
2	0.050	0.769											
	0.080	0.538	0.774										
	0.125	0.413	0.554	0.781									
	0.200	0.327	0.420	0.541	0.723								
	0.320	0.271	0.335	0.415	0.522	0.762							
	0.500	0.235	0.282	0.338	0.411	0.546	0.765						
	0.800	0.210	0.245	0.286	0.336	0.425	0.549	0.776					
	1.250		0.224	0.255	0.292	0.354	0.436	0.567	0.766				
	2.000			0.236	0.263	0.308	0.364	0.447	0.558	0.729			
	3.200					0.284	0.324	0.379	0.449	0.544	0.753		
	5.000							0.504	0.393	0.454	0.572	0.769	
	8.000								0.367	0.407	0.479	0.581	0.781
4	0.050	0.723											
	0.080	0.510	0.753										
	0.125	0.395	0.527	0.712									
	0.200	0.315	0.402	0.515	0.723								
	0.320	0.262	0.323	0.398	0.498	0.718							
	0.500	0.228	0.273	0.326	0.394	0.520	0.745						
	0.800	0.205	0.239	0.278	0.325	0.408	0.529	0.727					
	1.250		0.219	0.248	0.283	0.342	0.420	0.540	0.737				
	2.000			0.232	0.258	0.300	0.353	0.431	0.533	0.715			
	3.200					0.278	0.316	0.369	0.434	0.522	0.718		
	5.000							0.338	0.383	0.440	0.548	0.727	
	8.000								0.361	0.398	0.465	0.561	0.750

SAMPLING PLANS FOR AQL AND AOQL VARIABLES WITH MEASUREMENT ERROR

Table 3 (continued): Pa(pm) Values of Known Sigma Plans Under Measurement Error

r	AOQL (%)	AQL (%)											
		0.040	0.065	0.100	0.150	0.250	0.400	0.650	1.000	1.500	2.500	4.000	6.500
6	0.050	0.702											
	0.080	0.505	0.728										
	0.125	0.391	0.521	0.711									
	0.200	0.313	0.398	0.509	0.689								
	0.320	0.260	0.320	0.394	0.493	0.696							
	0.500	0.227	0.271	0.324	0.391	0.514	0.714						
	0.800	0.205	0.238	0.276	0.323	0.405	0.519	0.721					
	1.250		0.218	0.247	0.282	0.340	0.416	0.535	0.713				
	2.000			0.231	0.256	0.299	0.351	0.428	0.528	0.707			
	3.200					0.277	0.314	0.367	0.431	0.518	0.701		
	5.000							0.337	0.381	0.437	0.544	0.714	
	8.000								0.359	0.397	0.462	0.559	0.728
∞	0.050	0.700											
	0.080	0.501	0.727										
	0.125	0.389	0.515	0.700									
	0.200	0.311	0.395	0.505	0.663								
	0.320	0.258	0.318	0.392	0.489	0.696							
	0.500	0.226	0.270	0.322	0.389	0.510	0.714						
	0.800	0.203	0.236	0.274	0.321	0.402	0.514	0.719					
	1.250		0.227	0.246	0.281	0.338	0.413	0.530	0.702				
	2.000			0.104	0.255	0.297	0.349	0.425	0.524	0.672			
	3.200					0.276	0.313	0.365	0.428	0.514	0.696		
	5.000							0.336	0.379	0.435	0.540	0.714	
	8.000								0.358	0.395	0.460	0.553	0.720

Figure 1: Average Outgoing Quality Curves under Measurement Error

Greenberg, B. S., & Stokes, S. L. (1992). Estimating nonconformance rates after zero defect sampling with rectification. *Technometrics*, 34, 203-213.

Greenberg, B. S., & Stokes, S. L. (1995). Repetitive testing in the presence of inspection errors. *Technometrics*, 37, 102-111.

Hahn, G. J. (1986). Estimating the percent nonconforming in the accepted product after zero defect sampling. *Journal of Quality Technology*, 18, 182-188.

Hand, D. J. (2004). *Measurement theory and practice*. London, UK: Arnold.

SAMPLING PLANS FOR AQL AND AOQL VARIABLES WITH MEASUREMENT ERROR

Jackson, E. J. (1957). Effect of inspection errors on waste and on quality of outgoing product. *Industrial Quality Control*, *VIX*(6), 5-8.

Johnson, N. L., Kotz, S., & Wu, X. (1991). *Inspection errors for attributes in quality control*. London, UK: Chapman & Hall.

Martz, H. F., & Zimmer, W.J. (1990). A nonparametric Bayes empirical Bayes procedure for estimating the percent nonconforming in accepted lots. *Journal of Quality Technology*, *22*, 95-104.

Zaslavsky, A. (1988). Estimating nonconformity rates in C-defect sampling. *Journal of Quality Technology*, *20*, 248-259.

An Extension of Cochran-Orcutt Procedure for Generalized Linear Regression Models with Periodically Correlated Errors

Abdullah A. Smadi
Yarmouk University,
Irbid, Jordan

Nour H. Abu-Afouna
Nourah University,
Riyadh, Saudi Arabia

An important assumption of ordinary regression models is independence among errors. This research considers the case of periodically correlated errors following the periodic AR model of order 1 (PAR(1)). The remedial measure for correlated errors in regression known as the Cochran-Orcutt procedure is generalized to the case of periodically correlated errors. The motivation for making such generalizations is that the response data may inhibit some seasonality, which may not be captured by the traditional AR(1) autoregressive model. The proposed procedure is described and the bias and MSE of the resulting intercept and slope parameter estimates of the simple LR model with errors following PAR(1) are compared with those of ordinary least squares (OLS) estimates via simulation. An application of real data is provided.

Key words: Simple linear regression model, autoregression, periodic autoregression, Cochran-Orcutt procedure, autocorrelated errors.

Introduction

Assuming that $\{Y_1, Y_2, \dots, Y_n\}$ is an observed time series, then, using standard regression analysis suitable models of $\{Y_t\}$ may be developed. For example, if $\{Y_t\}$ consists of a deterministic trend along some random error, then $\{Y_t\}$ can be modeled as

$$Y_t = TR_t + \varepsilon_t$$

which contains, as a special case, the linear trend model

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t, t = 1, \dots, n \quad (1)$$

where the ε_t 's are usually assumed independent and identically distributed (iid) $N(0, \sigma_\varepsilon^2)$. This model can be generalized to other types of

trends, for example the polynomial trend. If, along with the trend, $\{Y_t\}$ also contains some deterministic seasonality, then extra terms are added to the trend model to capture seasonality.

The linear trend model in (1) is a special case of the simple linear regression (SLR) model

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \quad (2)$$

The inference of this model is straightforward. The ordinary least squares (OLS) estimators of β_0 and β_1 are given by

$$\left. \begin{aligned} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} \end{aligned} \right\} \quad (3)$$

and

$$\left. \begin{aligned} \text{Var}(\hat{\beta}_0) &= \frac{\sigma_\varepsilon^2}{S_{XX}} \left[\frac{S_{XX}}{n} + \bar{X}^2 \right] \\ \text{Var}(\hat{\beta}_1) &= \frac{\sigma_\varepsilon^2}{S_{XX}} \end{aligned} \right\} \quad (4)$$

Abdullah A. Smadi is an Associate Professor in the Department of Statistics. Email him at: asmadi@yu.edu.jo. Nour H. Abu-Afouna is an Instructor in the Department of Mathematics. Email her at: nouraboafouneh@yahoo.com.

where $S_{XX} = \sum (X_i - \bar{X})^2$ and $S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$. Under the assumptions of independence and constant variance, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the best linear unbiased estimators. In addition, they are maximum likelihood estimators under the normality assumption (Kutner, et al., 2005).

An important assumption of the model, which is frequently violated with time series data, is independence of errors $\{\epsilon_t\}$. Therefore, before adopting the OLS estimates data should be tested for independence among errors. If the assumption is not satisfied, then a remedial measure should be taken (Kutner, et al., 2005). In this article a remedial measure for regression models with correlated errors, namely the Cochran-Orcutt (COR) procedure is defined and generalized to the case of periodically correlated errors.

Testing for Correlated Errors

If the assumption of independence among errors in the regression model is violated, then the standard results about OLS estimators and their properties are questionable. An important diagnostic-checking method for this assumption is the Durbin-Watson (DW) test, which is commonly used, particularly when data are related to time as in (1). The DW test assumes a first-order autoregressive (AR(1)) model for errors, that is

$$\epsilon_t = \phi \epsilon_{t-1} + a_t \quad (5)$$

where u_t are assumed iid $N(0, \sigma_a^2)$ and $|\phi| < 1$. The DW test examines the presence of first order autocorrelation among errors ($\phi \neq 0$) against the null of white noise (WN) errors ($\phi = 0$). The most common version of this test is for positive autocorrelation ($\phi > 0$) (Bowerman, et al., 2005) and the test statistic is given by

$$D = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

where $e_t = Y_t - \hat{Y}_t$, $t = 1, \dots, n$ are the residuals of the OLS model (Kutner, et al., 2005, p. 487).

Although the DW test originally assumes an AR(1) model of errors, a significant result does not necessarily imply that the correct model of errors is AR(1) (Blattberg, 1973; Zinde-Walsh & Galbraith, 1991). In addition, many alternatives to the DW test are available, including the runs and the Breusch-Godfrey test (Breusch, 1979; Godfrey, 1978). Thus, if the DW test is found to be significant, the best model for the errors should be identified. In general, the errors model may be extended to the wider class of auto-regressive moving average (ARMA) models (Box, et al., 1994). Assuming that the errors are correlated and follow the AR(1) model, then suitable estimation methods are required; these include, but are not limited to the generalized least squares (GLS) method (see Lee & Lund, 2004) and the Cochran-Orcutt (COR) procedure.

Assuming that $\{Y_t\}$ (and possibly $\{X_t\}$) is a seasonal time series (TS) with period ω , the SLR model (2) can be rewritten as

$$Y_{k\omega+v} = \beta_0 + \beta_1 X_{k\omega+v} + \epsilon_{k\omega+v} \quad (6)$$

where $v = 1, \dots, \omega$ denotes the season and k denotes the year. In this case, if the errors in (1) are correlated, then they may inhibit some seasonality. In this case there several approaches exist to address the issue, which include adding terms to the regression model that capture seasonality as dummy variables, adding trigonometric functions or using seasonal ARMA models (Box, et al., 1994).

An alternative model is the periodic ARMA (PARMA) model (Tiao & Grupe, 1980; Franses & Paap, 2004). Writing the time t in terms of the period ω as $k\omega+v$, the special case of the zero-mean $PAR_\omega(1)$ model is

$$\epsilon_{k\omega+v} = \phi_1(v) \epsilon_{k\omega+v-1} + a_{k\omega+v}, \quad v = 1, \dots, \omega \quad (7)$$

where $\{a_{k\omega+v}\}$ is a zero-mean WN process with periodic variances $\sigma_a^2(v)$ and $\phi_1(v)$ is the AR parameter of season v . If the period $\omega = 1$, then this model reduces to the AR(1) model (5). It is assumed that this model is periodic stationary,

that is, $\left| \prod_{v=1}^{\omega} \phi_1(v) \right| < 1$ (Obeysekera & Salas, 1986). The properties of the OLS estimates of the SLR model when the errors are PAR(1) were investigated by Smadi and Abu-Afouna (2012). They also developed a GLS estimation for LR models under this setting of errors. PARMA models, which were first used in hydrology, are suitable for modeling periodic correlations; they have since become common in economic and other areas (Obeysekera & Salas, 1986; Franses & Paap, 2004).

The power of the DW test when errors are PAR(1) was investigated by Albertson, et al. (2002) who showed that the test is usually significant in this case. The DW test is, therefore, a good method to detect autocorrelations among errors, but it does not necessarily correctly identify its model (Lee & Lund, 2004).

An alternative test to the DW test was proposed by McLeod (1995). This test is designed for testing periodically autocorrelated errors. Assuming $n = m\omega$, then the residual $\{e_t\}$ is rewritten as $\{e_{k\omega+v}\}$ for $k = 0, \dots, m-1$ and $v = 1, \dots, \omega$; thus, the season-wise residuals can be obtained. For example, $v = 1$, $\{e_1, e_{\omega+1}, \dots, e_{(m-1)\omega+1}\}$ are the residuals for season 1, therefore, the first lag sample autocorrelation for season v is

$$r_1(v) = \frac{C_1(v)}{\sqrt{C_0(v)C_0(v-1)}},$$

where $C_0(v)$ and $C_1(v)$ are the sample variance for season v and the first lag sample seasonal autocovariance of season v , respectively, are given by

$$C_0(v) = \frac{\sum_{j=0}^{m-1} (e_{j\omega+v} - \bar{e}_v)^2}{m-1}$$

and

$$C_1(v) = \frac{\sum_{j=0}^{m-1} (e_{j\omega+v} - \bar{e}_v)(e_{j\omega+v-1} - \bar{e}_{v-1})}{m-1}.$$

McLeod demonstrated that $L = n \sum_{v=1}^{\omega} (r_1(v))^2$ is asymptotically distributed as a Chi-square with ω degrees of freedom under the assumption that there is no autocorrelation in the first lag for all seasons; thus, if $L > \chi_{\omega, \alpha}^2$ then it may be concluded that the errors are periodically autocorrelated. This test is implemented in R via the pear library (McLeod & Balcilar, 2008).

Generalization of Cochran-Orcutt Procedure for Errors Following PAR(1)

If error terms are autocorrelated, then the parameter estimation of the regression model is not straightforward. Assuming that the errors follow the AR(1) model (5), the SLR model in (2) is renamed as the generalized simple linear regression (GLR) model (Kutner, et al., 2005, p. 484). In this case, this model can be rewritten as (Kutner, et al., 2005, p. 491):

$$Y'_t = \beta'_0 + \beta'_1 X'_t + a_t, t = 1, \dots, n \quad (8)$$

where

$$\left. \begin{aligned} Y'_t &= Y_t - \rho Y_{t-1} \\ X'_t &= X_t - \rho X_{t-1} \end{aligned} \right\} \quad (9)$$

$$\left. \begin{aligned} \beta'_0 &= \beta_0(1-\rho) \\ \beta'_1 &= \beta_1 \end{aligned} \right\} \quad (10)$$

and

$$a_t = \varepsilon_t - \rho \varepsilon_{t-1}, t = 1, \dots, n \quad (11)$$

where $\{a_t\}$ is uncorrelated. Thus, (8) is a standard SLR model and the estimation of β_0 and β_1 begins by estimating ρ , then estimating β'_0 and β'_1 in (8) and finally obtaining estimates for β_0 and β_1 using (10). Several methods exist for estimating ρ in this situation, including the COR and Hildreth-Lu procedures (Kutner, et al., 2005). This study only considers the Cochran-Orcutt procedure, which involves an iteration of three steps (Kutner, et al., 2005, p. 492):

1. Estimation of ρ . This is accomplished by noting that the autoregressive error process assumed in model (2) can be viewed as a regression through the origin:

$$\varepsilon_t = \rho \varepsilon_{t-1} + a_t, t = 1, \dots, n.$$

Because ε_t and ε_{t-1} are unknown, residuals e_t and e_{t-1} obtained by OLS are used as the response and predictor variables, respectively, and ρ is estimated by fitting a straight line through the origin so that the moment estimator of the slope ρ is:

$$\hat{\rho} = \frac{\sum_{t=1}^n e_{t-1} e_t}{\sum_{t=2}^n e_{t-1}^2}. \quad (12)$$

2. Fitting of transformed model (8). Using the estimate $\hat{\rho}$ in (12), the transformed variables Y'_t and X'_t in (9) are obtained and OLS is used with these transformed variables to yield the fitted regression function as:

$$\hat{Y}'_t = \hat{\beta}'_0 + \hat{\beta}'_1 X'_t.$$

3. The DW test is employed to test whether the error terms for the transformed model are uncorrelated. If the test indicates that they are uncorrelated, the procedure terminates and $\hat{\beta}'_0$ and $\hat{\beta}'_1$ are obtained based on $\hat{\beta}'_0$ and $\hat{\beta}'_1$ in the Step 2 and by using (10).
4. If the DW test in Step 3 is significant, then steps (1)-(3) are repeated for Y' and X' in place of Y and X , and this continues until the DW test indicates that error terms are uncorrelated.
5. An estimate of σ_ε^2 is given by (Kutner, et al., 2005, p. 487) as:

$$\hat{\sigma}_\varepsilon^2 = \frac{\hat{\sigma}_a^2}{1 - \hat{\rho}^2}$$

where $\hat{\sigma}_a^2$ is the sample variance of residuals obtained from the fitted regression model in Step 2.

Consider the GLR model (2) with error terms following the zero-mean $\text{PAR}_\omega(1)$ model (7). The COR procedure described is now generalized to the GLR model with $\text{PAR}(1)$ errors. Assuming that Y_t and X_t are seasonal time series with period ω , (2) and (7) can be restated as:

$$\left. \begin{aligned} Y_{k,v} &= \beta_0 + \beta_1 X_{k,v} + \varepsilon_{k,v}, \\ \varepsilon_{k,v} &= \phi_1(v) \varepsilon_{k,v-1} + a_{k,v} \end{aligned} \right\} \quad (13)$$

where the time k denotes the year and $v = 1, 2, \dots, \omega$ denotes the season.

Theorem 1

The generalized regression model (13) is equivalent to:

$$Y'_{k,v} = \beta'_0(v) + \beta'_1(v) X'_{k,v} + a_{k,v}, \quad (14)$$

with

$$\left. \begin{aligned} Y'_{k,v} &= Y_{k,v} - \phi_1(v) Y_{k,v-1} \\ X'_{k,v} &= X_{k,v} - \phi_1(v) X_{k,v-1} \\ \beta'_0(v) &= \beta_0(1 - \phi_1(v)) \\ \beta'_1(v) &= \beta_1 \end{aligned} \right\}. \quad (15)$$

Proof 1

Substituting for $Y_{k,v}$ and $Y_{k,v-1}$ form (13) in $Y'_{k,v} = Y_{k,v} - \phi_1(v) Y_{k,v-1}$ gives

$$\begin{aligned} Y'_{k,v} &= (\beta_0 + \beta_1 X_{k,v} + \varepsilon_{k,v}) \\ &\quad - \phi_1(v) (\beta_0 + \beta_1 X_{k,v-1} + \varepsilon_{k,v-1}) \\ &= \beta_0(1 - \phi_1(v)) + \beta_1(X_{k,v} - \phi_1(v) X_{k,v-1}) \\ &\quad + (\varepsilon_{k,v} - \phi_1(v) \varepsilon_{k,v-1}) \end{aligned}$$

The transformed model in (14) is a GLR model with errors following a seasonal white noise process with periodic coefficients. To estimate the parameters of this model note that (14) defines a standard regression model for each season separately. That is, to estimate $\beta'_0(1)$, $\beta'_1(1)$ and $\sigma_a^2(1)$ only the data for $Y'_{k,1}$ and $X'_{k,1}$ is used. To summarize the generalized

COR procedure for errors following the PAR $\omega(1)$ model:

1. Using the OLS method, regress Y_t on X_t to obtain the residuals $\{e_t\}$.
2. Apply the DW test for autocorrelation among residuals; if residuals are not autocorrelated then the procedure terminates.
3. Estimate $\phi_1(v)$ by regressing $Y_{k,v}$ on $X_{k,v}$ for each season $v = 1, 2, \dots, \omega$ separately, then obtain the residual for each model $e_{k,v}^*$. Estimate $\phi_1(v)$ using:

$$\hat{\phi}_1(v) = \frac{\sum_{k=1}^m e_{k,v-1}^* e_{k,v}^*}{\sum_{k=1}^m (e_{k,v-1}^*)^2}. \quad (16)$$

4. Compute $Y'_{k,v}$ and $X'_{k,v}$ using (15) and the estimates in (16), then regress $Y'_{k,v}$ on $X'_{k,v}$ for data in each season v , separately. This results in $\beta'_0(v)$, $\beta'_1(v)$ and

$$\hat{\sigma}_a^2(v) = \frac{\sum_{k=1}^m (e'_{k,v})^2}{m-2}.$$

5. Apply the DW test on $\{e'_{k,v}\}$ for each season $v = 1, 2, \dots, \omega$. If none of the cases is significant then the procedure terminates. If, however, in some seasons the DW test is significant then the ordinary COR procedure is applied to those seasons until the DW test is found to be insignificant for all seasons
6. Using $\beta'_0(v)$, $\beta'_1(v)$ and (15) find $\hat{\beta}_0$ and $\hat{\beta}_1$, which are unbiased estimators of β_0 and β_1 denoted as $\hat{\beta}_{0v}$ and $\hat{\beta}_{1v}$, $v = 1, 2, \dots, \omega$.
7. Step 6 results in ω estimates of β_0 and β_1 , thus β_0 and β_1 may be estimated by the average of these estimates:

$$\tilde{\beta}_0 = \frac{1}{\omega} \sum_{v=1}^{\omega} \hat{\beta}_{0v} \quad \text{and} \quad (17)$$

$$\tilde{\beta}_1 = \frac{1}{\omega} \sum_{v=1}^{\omega} \hat{\beta}_{1v}$$

To estimate the variances of $\{e_t\}$, (7) results in

$$\sigma_\varepsilon^2(v) = \phi_1^2(v) \sigma_\varepsilon^2(v-1) + \sigma_a^2(v); v = 1, 2, \dots, \omega. \quad (18)$$

Replacing $\phi_1(v)$ and $\sigma_a^2(v)$ with the estimates obtained results in a system of ω equations that can be solved for $\sigma_\varepsilon^2(v)$.

It should be emphasized that the PAR models were chosen because they allow for periodic correlations between successive seasons that need not be homogeneous. Franses and Paap (2004) showed that many business time series data sets inhibit periodic autocorrelations. McLeod (1995) showed that the errors resulting from fitting seasonal ARMA models for several real-time series have periodic autocorrelations. Albertson and Aylen (1999) identified a PAR error process in modeling scrap steel market. Lastly, according to Osborn, et al. (1988), failure to allow for periodicity in time series data may bias specification tests and further complicate the modeling process.

OLS and COR Estimator Comparison

Estimates of β_0 and β_1 for the OLS method and the generalized COR method are next discussed and compared via bias and MSE using Monte-Carlo simulation; the focus is on the estimates of β_0 and β_1 only. For the simulation, an R-code was developed by the authors to run 2,000 repetitions each of realization length $4n$ ($n = 30, 50, 100$) for pairs of data (X, Y) . The simulation ran as follows:

1. Generate the predictor values $X_t = t + 2\cos(2\pi t/4)$, $t = 1, \dots, 4n$.
2. Generate the errors $\{\varepsilon_1, \varepsilon_1, \dots, \varepsilon_{n\omega}\}$ from the zero mean PAR $_4(1)$ model: $\varepsilon_{k,v} = \phi_1(v)\varepsilon_{k,v-1} + a_{k,v}$, with $\phi_1(1) = -0.9$,

$\phi_1(2) = 0.6$, $\phi_1(3) = 0.3$, $\phi_1(4) = -0.8$ and $\{a_{k,v}\}$ is a seasonal $WN-N(0, \sigma_a^2(v))$ with $\sigma_a^2(1) = 100$, $\sigma_a^2(2) = 1$, $\sigma_a^2(3) = 1$ and $\sigma_a^2(4) = 10$.

3. Compute $Y_t = 2 + 50X_t + \varepsilon_t$; $t = 1, \dots, 4n$.
4. Regress Y_t on X_t to obtain the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Apply Steps 1-7 of the generalized COR procedure and obtain $\tilde{\beta}_0$ and $\tilde{\beta}_1$ using (17).

Based on the realizations, the bias and MSE of estimates β_0 and β_1 for both the OLS and COR methods were computed and are presented in Table 1. The resulting OLS estimates are not reliable regardless of bias and MSE because the assumptions of the SLR model are not satisfied. Notice that the bias and MSE of estimates of β_0 and β_1 for both methods decrease as n increases. The proposed method estimates dominate the OLS estimates both in view of bias and MSE. Finally, the differences in bias and MSE for both methods were more apparent for the estimates of β_0 compared to those for β_1 .

Table 1: Bias and MSE (in brackets) of β_0 and β_1 Estimates

n	OLS		COR	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\tilde{\beta}_0$	$\tilde{\beta}_1$
30	0.0003 (0.0006)	0.0206 (3.0510)	-0.0055 (0.6821)	-0.000005 (0.0001)
50	-0.0003 (0.0360)	0.0360 (1.8351)	-0.0163 (0.3137)	0.00008 (0.00002)
100	-0.0001 (0.00001)	0.0076 (0.9399)	-0.0049 (0.1389)	0.000003 (0.000002)

Application to Real Data

Consider quarterly U.S. airline passenger-miles (in millions). The data (which was originally monthly but was aggregated to

quarterly) shows 9 years from 1996 to 2004 (Cryer & Chan, 2008). Figure 1 shows the data, denoted by Y_t ; $t = 1, \dots, 36$. The time series shows both a nearly increasing trend and an apparent seasonality. This data set is used to illustrate the proposed method as discussed previously. The generalized COR procedure was applied as follows:

1. The linear trend model was fitted for Y_t and $\hat{Y}_t = 104.608 + 0.866t$. Assuming the errors are WN, the estimated error variance is the MSE of the OLS regression, that is, $\hat{\sigma}_e^2 = 2771.9$.
2. Based on the residuals $\{e_t\}$ in Step 1 the DW test was applied and resulted in a significant p-value of 0.003. To check that the errors are periodically autocorrelated the McLeod test was also applied on $\{e_t\}$ with period $\omega = 4$. The p-value equals 0.00009, which is also highly significant; this indicates that there is sufficient evidence to suggest that the errors are periodically autocorrelated. Figure 2 shows the variability among residuals for various quarters; the ACF shows significant correlations at lag one, which agrees with the DW test, and is also significant at lag four, which is the seasonal lag.
3. Y_t and t are subdivided by quarters. For each $v = 1, \dots, 4$, $Y_{k,v}$ is regressed on $t_{k,v}$, $k = 1, \dots, 9$. The four fitted regression models were:

$$\hat{Y}_{k,1} = 99.544 + 0.733t_{k,1}$$

$$\hat{Y}_{k,2} = 108.495 + 0.925t_{k,2}$$

$$\hat{Y}_{k,3} = 110.877 + 0.914t_{k,3}$$

and

$$\hat{Y}_{k,4} = 100.211 + 0.847t_{k,4}$$

thus, $\hat{\phi}_1(1) = 0.733$, $\hat{\phi}_2(2) = 0.925$, $\hat{\phi}_3(3) = 0.914$ and $\hat{\phi}_4(4) = 0.847$.

Figure 1: Time Series Plot of Quarterly U.S. Airline Passenger Miles, 1996-2004

Figure 2: The Parallel Box Plot of Residuals (top) and ACF of Residuals (bottom) for the fitted OLS Model

4. $Y'_{k,v}$ and $t'_{k,v}$ were obtained using (15). Results obtained when regressing $Y'_{k,v}$ on $t'_{k,v}$ for each quarter separately (see Table 2).

Table 2: Estimates of Transformed Regressions for Various Quarters

v	$\hat{\beta}'_0(v)$	$\hat{\beta}'_1(v)$	$\hat{\sigma}_a^2(v)$
1	23.081	1.081	18.045
2	14.339	3.190	4.135
3	10.527	1.210	18.181
4	6.0950	0.584	22.482

5. The residuals $\{e'_{k,v}\}$ for each season $v = 1, \dots, 4$ were computed from the fitted models in Step 4 and the DW test is applied for each season. It was found that p-values for all tests were: 0.481, 0.317, 0.273 and 0.419, thus, not all are significant so that the iterations terminate.
6. $\hat{\beta}_{0v}$ and $\hat{\beta}_{1v}$ were obtained as shown in Step 6 of the proposed method described previously and are based on (15). Each season v separately gave:

$$\begin{aligned}\hat{\beta}_{01} &= 86.468, & \hat{\beta}_{11} &= 1.081 \\ \hat{\beta}_{02} &= 190.195, & \hat{\beta}_{12} &= 3.190 \\ \hat{\beta}_{03} &= 122.848, & \hat{\beta}_{13} &= 1.210 \\ \hat{\beta}_{04} &= 39.937, & \hat{\beta}_{14} &= 0.584\end{aligned}$$

Using these estimates and (17) results in $\tilde{\beta}_0 = 109.862$ and $\tilde{\beta}_1 = 1.516$.

7. Estimates of the variances of $\{\epsilon_t\}$ were obtained using (18) and equal

$$\begin{aligned}\hat{\sigma}_\epsilon^2(1) &= 53.128, \\ \hat{\sigma}_\epsilon^2(2) &= 49.555, \\ \hat{\sigma}_\epsilon^2(3) &= 59.607, \\ \hat{\sigma}_\epsilon^2(4) &= 65.284.\end{aligned}$$

Note that the resulting estimate of intercept in Step 6 is very close to that of the OLS estimate in Step 1, however, a larger difference is detected in the estimate of the slope parameter. This is due to the fact that the slope is directly affected by the periodic correlations. The largest effect was observed on the estimate of the error terms, which was very large assuming WN errors (see Step 1) compared to estimates that account for periodic correlations in Step 7.

Finally, it should be noted that the objective of this application is for illustration of the proposed method. The magnitudes of differences between OLS and the proposed method estimates do not necessarily count for our method, meaning that, after the errors are correlated the OLS estimates and their standard errors are not reliable. Because this article focuses on the fact that in standard regression analysis, particularly when dealing with time series data, routine residual analysis should test for autocorrelation among errors and determine whether it is a traditional AR(1) autocorrelation or periodic.

Conclusion

This study examined the SLR model with correlated errors. The ordinary Cochran-Orcutt procedure for SLR models with correlated errors with AR(1) model was generalized to the case of periodically correlated errors as a PAR(1) model, which produced estimates of regression parameters β_0 and β_1 . Monte Carlo simulations were used to compare the ability of both methods to estimate β_0 and β_1 via bias and MSE. Results indicate that the estimates based on the proposed COR procedure dominate the OLS estimates.

This study was designed to consider the fact that errors in ordinary regression analysis may exhibit periodic autocorrelation which can be modeled by a PAR(1) model and not

necessarily an AR(1) model. The proposed method is by no means the best remedial measure for periodic correlations. Future research may add seasonal dummies to the regression model or to use generalized least squares regression.

References

- Albertson, K., & Aylen, J. (1999). Forecasting using a periodic transfer function with an application to the UK price of ferrous scrap. *International Journal of Forecasting*, 15, 409-419.
- Albertson, K., Aylen, J. & Lim, K. B. (2002). The power of the Durbin Watson test when the errors are PAR(1). *Journal of Statistical Computation and Simulation*, 72, 507-516.
- Blattberg, R. (1973). Evaluation of the power of the Durbin-Watson statistic for non-first order serial correlation alternatives. *Journal Review of Economics & Statistics*, 55, 508-515.
- Breusch, T. (1979). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, 17, 334-355.
- Box, G., Jenkins, G., & Reinsel, G. (1994). *Time series analysis, forecasting and control*, 3rd Ed. Englewood Cliffs, NJ: Prentice-Hall.
- Bowerman, B., O'Connell, R., & Koehler, A. (2005). *Forecasting, time series and regression, an applied approach*, 4th Ed. Belmont, CA.:Thomson Brooks/Cole.
- Cryer, J. D., & Chan, K. (2008). *Time series analysis with applications in R*, 2nd Ed. New York: Springer-Verlag.
- Franses, P., & Paap, R. (2004). *Periodic time series models*. Oxford, UK: Oxford University Press.
- Godfrey, L. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, 46, 1293-1302.
- Kutner, H., Nachtsheim, J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models*, 5th Ed. LONDON: McGraw-Hill.
- Lee, J., & Lund, R. (2004). Revisiting simple linear regression with autocorrelated errors. *Biometrika*, 91, 240-245.
- McLeod, A. I. (1995). Diagnostic checking periodic autoregression models with application. *Journal of Time Series Analysis*, 15, 221-233.
- McLeod, A. I., & Balcilar, M. (2008). *Pear: Package for periodic autoregression analysis*. R package version 1.0.,<http://www.r-project.org>.
- Obeysekera, J. T. B., & Salas, J. D. (1986). Modeling of aggregated hydrologic time series. *Journal of Hydrology*, 86, 197-219.
- Osborn, D., Chui, A., Smith, J., & Birchenhall, C. (1988). Seasonality and the order of integration for consumption. *Oxford Bulletin of Economics and Statistics*, 50, 361-377.
- Smadi, A., & Abu-Afouna, N. (2012). On the least squares estimation of simple linear regression model with periodically correlated errors: a cautionary note. Accepted in *Austrian Journal of Statistics*.
- Tiao, G., & Grupe, M. (1980). Hidden periodic autoregressive-moving average models in time series data. *Biometrika*, 67, 365-373.
- Zinde-Walsh, V., & Galbraith, J. (1991). Estimators of a linear regression model with stationary ARMA(p, q) errors. *Journal of Econometrics*, 47, 333-357.

Bayesian Estimation of Erlang Distribution under Different Generalized Truncated Distributions as Priors

Adil H. Khan T. R. Jan
University of Kashmir,
Srinagar. India

Various generalized truncated distributions are considered as independent informative priors for estimating shape and scale parameters of the Erlang distribution. In addition, various special cases are also discussed.

Key words: Erlang distribution, generalized truncated distributions, Bayes estimator, posterior distribution.

Introduction

The Erlang distribution is a continuous probability distribution with wide applicability, primarily due to its relation to the exponential and Gamma distributions. The Erlang distribution was developed by A. K. Erlang (1909) to examine the number of telephone calls that could be made at the same time to switching station operators. This work on telephone traffic engineering has been expanded to consider waiting times in queuing systems in general. Queuing theory originated when Erlang (1909) published his fundamental paper relating to the study of telephone traffic congestion (Brockmeyer, Halstorm & Jensen, 1948).

The probability function of the Erlang distribution is

$$f(x : u, v) = \frac{x^{u-1} \exp(-v^{-1}x)}{\Gamma(u) v^u}; \quad (1.1)$$
$$u = 1, 2, 3, \dots; v > 0, x > 0$$

where u and v are unknown parameters and are respectively called shape and scale parameters. When the scale parameter $v = 2$, the distribution simplifies to a Chi-squared distribution with $2k$ degrees of freedom; therefore, it can be regarded as a generalized Chi-squared distribution.

The Erlang distribution is the distribution of u independent identically distributed random variables each with an exponential distribution, and can be expressed as waiting time and message length in telephone traffic. If the durations of individual calls are exponentially distributed, then the duration of successive calls is the Erlang distribution. The Erlang distribution is a special case of the Gamma distribution when the shape parameter u is an integer (Evans, et al., 2000).

Harischandra and Rao (1998) discussed problems with classical inferences for the Erlangian queue. Bhattacharyya and Singh (1994) obtained a Bayes estimator for the Erlangian queue under two prior densities. Wiper (1998) studied $Er/M/1$ and $Er/M/C$ queues under a Bayesian setup and estimated equilibrium probabilities of queue size and waiting time distribution using conditional Monte-Carlo simulation methods. Jain (2001) examined the problem of change point for the inter-arrival time distribution in the context of exponential families for the $Ek/G/e$ queuing system and obtained a Bayes estimate of posterior probabilities and the position of the change from the Erlang distribution. Nair, et al.

Adil H. Khan is a Research Scholar in the Post-Graduate Department of Statistics. Email him at: khanadil_192@yahoo.com. T. R. Jan is an Assistant Professor in the Post-Graduate Department of Statistics, University of Kashmir. Email him at: drtrjan@gmail.com.

(2003) studied the Erlang distribution as a model for ocean wave periods and obtained different characteristics of the distribution under a classical setup. Suri, et al (2009) used the Erlang distribution to design a simulator for project management process time estimation. Recently, Damodaran, et al. (2010) obtained the expected time between failure measures and showed that the predicted failure times are closer to the actual failure time. Haq and Dey (2011) addressed the problem of Bayesian estimation of parameters for the Erlang distribution assuming different independent informative priors. This article estimates parameters of the Erlang distribution using different generalized truncated distributions.

Prior and Loss Functions

In many practical situations information is available regarding the shape and scale parameters of sampling distributions, therefore, this article considers a number of prior distributions and assumes that the parameters u and v are independent. The distributions considered, which are priors for shape and scale parameters, are:

- Generalized Truncated Poisson;
- Generalized Truncated Geometric;
- Generalized Truncated Poisson and Inverted Gamma;
- Generalized Truncated Poisson and Gamma;
- Generalized Truncated Geometric and Inverted Gamma; and the
- Generalized Truncated Geometric and Gamma.

The loss function considered is a squared error loss function. The squared error loss function for the shape parameters (u) and the scale parameters (v) are defined as

$$L(\hat{u}) = (\hat{u} - u)^2$$

and

$$L(\hat{v}) = (\hat{v} - v)^2,$$

which are symmetric, and where u and v , and $\hat{u}$ and $\hat{v}$ represent the true and estimated values of the parameters.

Derivation of Posterior Distribution under Different Informative Priors

If $X_1, X_2, X_3, \dots, X_n$ are a random sample from the Erlang distribution, then the likelihood function of sample observations $x_1, x_2, x_3, \dots, x_n$ is defined as:

$$L(u, v : x) = \frac{\prod_{i=1}^n x_i \exp(-v \sum_{i=1}^n x_i)}{(\Gamma(u))^n v^{nu}};$$

$$u = 1, 2, 3, \dots; v > 0$$
(3.1)

When shape parameter u is unknown and scale parameter v is known, then the performance of the Bayes estimators depends on the form of the prior distribution and the loss function assumed. Two different informative prior distributions are assumed for the shape parameter u , namely, the generalized truncated Poisson distribution and the generalized truncated geometric distribution. These are used to obtain the Bayes estimators and posterior variances. Also, Bayes estimators and posterior variances are also obtained for the truncated Poisson distribution and the truncated geometric distribution as the special cases.

Generalized Truncated Poisson Distribution as a Prior for Shape Parameter u

The probability density function (pdf) of the generalized truncated Poisson distribution is:

$$g_1(u_1, \theta_1) = \frac{(1 + \alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1 + \alpha u))}{\Gamma(u+1)(1 - \exp(-\theta_1))};$$

$$u = 1, 2, 3, \dots, \theta_1 > 0.$$
(3.1.1.1)

For $\alpha = 0$ the truncated Poisson distribution is:

$$g_1^*(u_1, \theta_1) = \frac{\theta_1^u \exp(-\theta_1)}{\Gamma(u+1)(1-\exp(-\theta_1))};$$

$$u = 1, 2, 3, \dots, \theta_1 > 0.$$

(3.1.1.2)

By combining likelihood function (3.1) with prior density function (4.1) the posterior distribution of u , the prior is the generalized truncated Poisson distribution:

$$g_1\left(\frac{u}{X}\right) = \frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}}; \\ \sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right);$$

$$u = 1, 2, \dots$$

(3.1.1.3)

Under the squared error loss function with the prior $g_1(u_1, \theta_1)$, the Bayes estimator is

$$\hat{u}_1 / X = \frac{\sum_{u=1}^{\infty} \left(\frac{u(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right)}.$$

(3.1.1.4)

The posterior variance of the Bayes estimator $\hat{u}_1 / X$ is given by:

$$\text{Var}\left(\frac{\hat{u}_1}{X}\right) = \frac{\left(\sum_{u=1}^{\infty} \left(\frac{u^2 (1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right) \right)}{\left(\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right) \right)} - \left(\frac{\sum_{u=1}^{\infty} \left(\frac{u(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right)} \right)^2$$

(3.1.1.5)

Special Cases

For $\alpha = 0$, the generalized truncated Poisson distribution reduces to the truncated Poisson distribution; therefore, the Bayes estimator for scale parameter u is given by

$$\hat{u}_1^* / X = \frac{\sum_{u=1}^{\infty} \left(\frac{u \theta_1^u \exp(-\theta_1) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\theta_1^u \exp(-\theta_1) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right)},$$

(3.1.1.6)

and the posterior variance of the Bayes estimator $\hat{u}_1^*/X$ is given by

$$\text{Var}\left(\hat{u}_1^*/X\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{u^2 \theta_1^u \exp(-\theta_1) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\theta_1^u \exp(-\theta_1) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right)} - \left[\frac{\sum_{u=1}^{\infty} \left(\frac{u \theta_1^u \exp(-\theta_1) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\theta_1^u \exp(-\theta_1) \exp(u \sum_{i=1}^n \ln x_i)}{\Gamma(u+1)(\Gamma(u))^n v^{nu}} \right)} \right]^2. \quad (3.1.1.7)$$

Generalized Truncated Geometric Distribution as a Prior for Shape Parameter u

If the prior assumed for the shape parameter is the Generalized Truncated Geometric distribution, then the pdf of the Generalized Truncated Geometric distribution is given by

$$g_2(u, \theta_2) = (1-m+\theta_2) \binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u}; \\ u=1, 2, 3, \dots; 0 < \theta_2 < 1, \quad (3.1.2.1)$$

and, for $m=1$, the truncated geometric distribution with pdf results as

$$g_2^*(u, \theta_2) = \theta_2 (1-\theta_2)^{u-1}; \\ u=1, 2, 3, \dots; 0 < \theta_2 < 1.$$

By combining likelihood function (3.1) and prior density function (5.1), the posterior distribution of u , when the prior is the generalized truncated geometric distribution, is

$$g_2(u/X) = \frac{\binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp(u \sum_{i=1}^n \ln x_i)}{(\Gamma(u))^n v^{nu}} \cdot \frac{1}{\sum_{u=1}^{\infty} \left(\frac{\binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp(u \sum_{i=1}^n \ln x_i)}{(\Gamma(u))^n v^{nu}} \right)}, \\ u=1, 2, 3, \dots \quad (3.1.2.2)$$

Under a squared error loss function with prior $g_2(u, \theta_2)$, the Bayes estimator is

$$\hat{u}_2^*/X = \frac{\sum_{u=1}^{\infty} \left(\frac{u \binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp(u \sum_{i=1}^n \ln x_i)}{(\Gamma(u))^n v^{nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp(u \sum_{i=1}^n \ln x_i)}{(\Gamma(u))^n v^{nu}} \right)}, \quad (3.1.2.3)$$

and the posterior variance of the Bayes estimator $\hat{u}_2^*/X$ is given by

$$\text{Var}\left(\hat{u}_2 \Big/ X\right) = \frac{\sum_{u=1}^{\infty} \left[\frac{u^2 \binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right)}{(\Gamma(u))^n v^{nu}} \right]}{\sum_{u=1}^{\infty} \left[\frac{\binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right)}{(\Gamma(u))^n v^{nu}} \right]} - \left[\frac{\sum_{u=1}^{\infty} \left[\frac{u \binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right)}{(\Gamma(u))^n v^{nu}} \right]}{\sum_{u=1}^{\infty} \left[\frac{\binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right)}{(\Gamma(u))^n v^{nu}} \right]} \right]^2. \quad (3.1.2.4)$$

Special Cases

For $m = 1$, the generalized truncated geometric distribution reduces to a truncated geometric distribution; therefore, the Bayes estimator for scale parameter u is given by:

$$\hat{u}_2^* \Big/ X = \frac{\sum_{u=1}^{\infty} \left[\frac{u (1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right)}{(\Gamma(u))^n v^{nu}} \right]}{\sum_{u=1}^{\infty} \left[\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right)}{(\Gamma(u))^n v^{nu}} \right]}, \quad (3.1.2.5)$$

and the posterior variance of Bayes estimator

$\hat{u}_2^* \Big/ X$ is given by

$$\text{Var}\left(\hat{u}_2^* \Big/ X\right) = \frac{\sum_{u=1}^{\infty} \left[\frac{u^2 (1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right)}{(\Gamma(u))^n v^{nu}} \right]}{\sum_{u=1}^{\infty} \left[\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right)}{(\Gamma(u))^n v^{nu}} \right]} - \left[\frac{\sum_{u=1}^{\infty} \left[\frac{u (1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right)}{(\Gamma(u))^n v^{nu}} \right]}{\sum_{u=1}^{\infty} \left[\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right)}{(\Gamma(u))^n v^{nu}} \right]} \right]^2. \quad (3.1.2.6)$$

When both the shape and scale parameters are unknown, the different independent prior distributions are assumed for two unknown parameters u (shape) and v (scale) of Erlang distributions.

Posterior Distribution under Generalized Truncated Poisson and Inverted Gamma Priors

The assumed prior for the shape parameter u of the Erlang distribution is Generalized Truncated Poisson distribution having the pdf

$$g_{11}(u, \theta_1) = \frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u))}{\Gamma(u+1)(1-\exp(-\theta_1))};$$

$$u = 1, 2, 3, \dots; \theta_1 > 0. \quad (3.2.1.1)$$

For the scale parameter v , the assumed prior is the inverted Gamma distribution with pdf

$$g_{12}(v; \alpha_1, \beta_1) = \frac{\beta_1^{\alpha_1} v^{-(\alpha_1+1)} \exp(-v^{-1}\beta_1)}{\Gamma(\alpha_1)};$$

$$v > 1, \alpha_1 > 0, \beta_1 > 0. \quad (3.2.1.2)$$

The joint prior distribution of u and v is defined as:

$$g_1(u, v) = g_{11}(u, \theta_1) \cdot g_{12}(v; \alpha_1, \beta_1). \quad (3.2.1.3)$$

By combining likelihood function (3.1) and joint prior function (6.3), the joint posterior

distribution of u and v is given by (3.2.1.4). The marginal posterior distribution of u and v , $g_1\left(u, \frac{v}{X}\right)$, are shown in (3.2.1.5) and (3.2.1.6).

Under the squared loss function, the expression for Bayes estimators of u and v with their respective posterior variances are given by (3.2.1.7), (3.2.1.8), (3.2.1.9) and (3.2.1.10).

$$\begin{aligned} g_1\left(u, \frac{v}{X}\right) &= \frac{\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \times \exp\left(u \sum_{i=1}^n \ln x_i\right) v^{-(\alpha_1+nu+1)} \times \exp\left\{-v^{-1}(\beta_1 + \sum_{i=1}^n x_i)\right\}}{\Gamma(u+1)(\Gamma(u))^n}}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \times \exp\left(u \sum_{i=1}^n \ln x_i\right) \int_0^{\infty} v^{-(\alpha_1+nu+1)} \times \exp\left\{-v^{-1}(\beta_1 + \sum_{i=1}^n x_i)\right\} db}{\Gamma(u+1)(\Gamma(u))^n} \right)} \\ &= \frac{\left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \times \exp\left(u \sum_{i=1}^n \ln x_i\right) v^{-(\alpha_1+nu+1)} \times \exp\left\{-v^{-1}(\beta_1 + \sum_{i=1}^n x_i)\right\}}{\Gamma(u+1)(\Gamma(u))^n} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \times \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1+nu}} \right)} \end{aligned} \quad (3.2.1.4)$$

$$\begin{aligned} g_1\left(\frac{u}{X}\right) &= \frac{\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \int_0^{\infty} v^{-(\alpha_1+nu+1)} \exp\left\{-v^{-1}(\beta_1 + \sum_{i=1}^n x_i)\right\} db}{\Gamma(u+1)(\Gamma(u))^n}}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1+nu}} \right)} \\ &= \frac{\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1+nu}}}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1+nu}} \right)}; \\ u &= 1, 2, \dots \end{aligned} \quad (3.2.1.5)$$

$$g_1\left(\frac{v}{X}\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) v^{-nu}}{\Gamma(u+1)(\Gamma(u))^n} \left(\frac{\exp\left\{-v^{-1}(\beta_1 + \sum_{i=1}^n x_i)\right\}}{v^{\alpha_1+1}} \right) \right)}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)},$$

(3.2.1.6)

$$\hat{u}_1/X = \frac{\sum_{u=1}^{\infty} \left(\frac{u(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)}$$

(3.2.1.7)

$$\begin{aligned} \hat{v}_1/X &= \frac{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right)}{\Gamma(u+1)(\Gamma(u))^n} \right) \left(\int_0^{\infty} \frac{v \exp\left\{-v^{-1}(\beta_1 + \sum_{i=1}^n x_i)\right\}}{v^{\alpha_1 + nu + 1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)} \\ &= \frac{\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 1)}{\Gamma(u+1)(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu - 1}}}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)} \end{aligned}$$

(3.2.1.8)

$$\text{Var}\left(\hat{u}_1/X\right) = \left[\frac{\sum_{u=1}^{\infty} \left(\frac{u^2 (1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)} \right] - \left[\frac{\sum_{u=1}^{\infty} \left(\frac{u(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)} \right]^2 \quad (3.2.1.9)$$

$$\text{Var}\left(\hat{v}_1/X\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right)}{\Gamma(u+1)(\Gamma(u))^n} \int_0^{\infty} \frac{v \exp\left\{-v^{-1} \left(\beta_1 + \sum_{i=1}^n x_i\right)\right\}}{v^{\alpha_1 + nu + 1}} dv \right)}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)} - \left[\frac{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 1)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu - 1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)} \right]^2 \quad (3.2.1.10)$$

$$\begin{aligned} \text{Var}\left(\hat{v}_1/X\right) = & \frac{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 2)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu - 2}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)} \\ & - \left[\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 1)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu - 1}} \right]^2 \\ & \left[\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right) \right] \end{aligned} \quad (3.2.1.10)$$

Special Cases

For $\alpha = 0$, the generalized truncated Poisson distribution reduces to the truncated Poisson distribution, therefore, the Bayes estimator for scale parameter u and shape parameter v are given by

$$\begin{aligned} \hat{u}_1^*/X = & \frac{\sum_{u=1}^{\infty} \left(\frac{u \theta_1^u \exp(-\theta_1) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\theta_1^u \exp(-\theta_1) \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)} \end{aligned} \quad (3.2.1.11)$$

and

$$\begin{aligned} \hat{v}_1^*/X = & \frac{\theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 1)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu - 1}} \\ & \sum_{u=1}^{\infty} \left(\frac{\theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right) \end{aligned} \quad (3.2.1.12)$$

In addition, the respective posterior variances of u and v are

$$\text{Var}\left(\hat{u}_1^*/X\right) = \left(\frac{\sum_{u=1}^{\infty} \left(\frac{u^2 \theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)} \right) - \left(\frac{\sum_{u=1}^{\infty} \left(\frac{u \theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)} \right)^2 \quad (3.2.1.13)$$

and

$$\text{Var}\left(\hat{v}_1^*/X\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{\theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 2)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu - 2}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)} - \left(\frac{\sum_{u=1}^{\infty} \left(\frac{\theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 1)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu - 1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{\Gamma(u+1)(\Gamma(u))^n \left(\beta_1 + \sum_{i=1}^n x_i\right)^{\alpha_1 + nu}} \right)} \right)^2 \quad (3.2.1.14)$$

Posterior Distribution under Generalized Truncated Poisson and Gamma Priors

If the prior for u is assumed to be a generalized truncated Poisson distribution and the assumed prior for v is a Gamma distribution with pdfs:

$$g_{11}(u_1, \theta_1) = \frac{(1 + \alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1 + \alpha u))}{\Gamma(u+1)(1 - \exp(-\theta_1))};$$

$$u = 1, 2, 3, \dots; \theta_1 > 0, \quad (3.2.2.1)$$

and

$$g_{21}(v; \alpha_2, \beta_2) = \frac{\beta_2^{\alpha_2} v^{-(\alpha_2+1)} \exp(-v^{-1}\beta_2)}{\Gamma(\alpha_2)};$$

$$v > 1, \alpha_2 > 0, \beta_2 > 0, \quad (3.2.2.2.)$$

then the joint Prior Distribution of u and v is defined as:

$$g(u, v) = g_{11}(u_1, \theta_1) \cdot g_{21}(v; \alpha_2, \beta_2) \quad (3.2.2.3)$$

Combining likelihood function (3.1) with joint density (7.3), the joint posterior distributions of u and v are given by (3.2.2.4). The marginal posterior distribution of u and v are given by (3.2.2.5) and (3.2.2.6). The expression for Bayes estimators of u and v with their respective posterior variance are given in (3.2.2.7), (3.2.2.8), (3.2.2.9) and (3.2.2.10).

$$\begin{aligned}
 & \frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \int_0^\infty v^{(\alpha_2-nu)-1} \exp\left\{\frac{-1}{2}\left((2\beta_2)v + \frac{(2\sum_{i=1}^n x_i)}{v}\right)\right\} db}{\exp\left(-u \sum_{i=1}^n \ln x_i\right) \Gamma(u+1)(\Gamma(u))^n} \\
 g_2\left(\frac{u}{X}\right) = & \frac{\sum_{u=1}^\infty \left[\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_2\left(2\sqrt{\beta_2(\sum x_i)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right]}{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu}\left(2\sqrt{\beta_2(\sum x_i)}\right)} \\
 & \frac{\Gamma(u+1)(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}}{\Gamma(u+1)(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}}; \\
 g_2\left(\frac{u}{X}\right) = & \frac{\sum_{u=1}^\infty \left[\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha-nu}\left(2\sqrt{\beta_2(\sum x_i)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right]}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}}; \\
 & u = 1, 2, \dots
 \end{aligned} \tag{3.2.2.5}$$

$$\begin{aligned}
 & \sum_{u=1}^\infty \left[\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) v^{(\alpha_2-nu)-1} \exp\left\{\frac{-1}{2}\left((2\beta_2)v + \frac{(2\sum_{i=1}^n x_i)}{v}\right)\right\}}{\exp\left(-u \sum_{i=1}^n \ln x_i\right) \Gamma(u+1)(\Gamma(u))^n} \right] \\
 g_2\left(\frac{v}{X}\right) = & \frac{\sum_{u=1}^\infty \left[\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu}\left(2\sqrt{\beta_2(\sum x_i)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right]}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}}; \\
 & v > 0
 \end{aligned} \tag{3.2.2.6}$$

$$\hat{u}_2/X = \frac{\sum_{u=1}^{\infty} \left(\frac{2u(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)}$$

(3.2.2.7)

$$\hat{v}_2/X = \frac{\sum_{u=1}^{\infty} \left(\frac{(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \int_0^{\infty} v \cdot v^{(\alpha_2-nu)-1} \exp\left\{\frac{-1}{2} \left((2\beta_2)v + \frac{(2\sum_{i=1}^n x_i)}{v} \right)\right\} db}{\exp\left(-u \sum_{i=1}^n \ln x_i\right) \Gamma(u+1)(\Gamma(u))^n} \right)}{\sum_{u=1}^{\infty} \left(\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)}$$

$$\hat{v}_2/X = \frac{\sum_{u=1}^{\infty} \left(\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu+1} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu+1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1(1+\alpha u)) \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)}$$

(3.2.2.8)

$$\text{Var}\left(\hat{u}_2/X\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{2u^2 (1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 (\sum x_i)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha-nu} \left(2\sqrt{\beta_2 (\sum x_i)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)}$$

$$- \left(\frac{\sum_{u=1}^{\infty} \left(\frac{2u(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 (\sum x_i)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha-nu} \left(2\sqrt{\beta_2 (\sum x_i)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)} \right)^2 \quad (3.2.2.9)$$

$$\text{Var}\left(\hat{v}_2/X\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu+2} \left(2\sqrt{\beta_2 (\sum x_i)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu+2}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha-nu} \left(2\sqrt{\beta_2 (\sum x_i)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)}$$

$$- \left(\frac{\sum_{u=1}^{\infty} \left(\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu+1} \left(2\sqrt{\beta_2 (\sum x_i)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu+1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2(1+\alpha u)^{u-1} \theta_1^u \exp(-\theta_1 (1+\alpha u)) \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha-nu} \left(2\sqrt{\beta_2 (\sum x_i)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{(\sum x_i)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)} \right)^2 \quad (3.2.2.10)$$

Special Cases

For $\alpha = 0$, the generalized truncated Poisson distribution reduces to the truncated Poisson distribution, therefore, the Bayes

estimator for scale parameter u and shape parameter v are given by (3.2.2.11) and (3.2.2.12). Their respective posterior variances are given by (3.2.2.13) and (3.2.2.14).

$$\hat{u}_2^*/X = \frac{\sum_{u=1}^{\infty} \left(\frac{2u \theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)} \quad (3.2.2.11)$$

$$\hat{v}_2^*/X = \frac{\sum_{u=1}^{\infty} \left(\frac{2 \theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu+1} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2-nu+1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \theta_1^u \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)} \quad (3.2.2.12)$$

$$\text{Var}\left(\hat{u}_2^*/X\right) = \left(\frac{\sum_{u=1}^{\infty} \left(\frac{2u^2 \theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)} \right) - \left(\frac{\sum_{u=1}^{\infty} \left(\frac{2u \theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2-nu}} \right)} \right)^2 \quad (3.2.2.13)$$

$$\text{Var}\left(\hat{v}_2^*/X\right) = \left(\frac{\sum_{u=1}^{\infty} \left(\frac{2 \theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2 - nu + 2}\left(2\sqrt{\beta_2\left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2 - nu + 2}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \theta_1^u \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha - nu}\left(2\sqrt{\beta_2\left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2 - nu}} \right)} \right) - \left(\frac{\sum_{u=1}^{\infty} \left(\frac{2 \theta_1^u \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2 - nu + 1}\left(2\sqrt{\beta_2\left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2 - nu + 1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \theta_1^u \exp\left(-u \sum_{i=1}^n \ln x_i\right) K_{\alpha - nu}\left(2\sqrt{\beta_2\left(\sum x_i\right)}\right)}{\Gamma(u+1)(\Gamma(u))^n \left(\sqrt{\left(\sum x_i\right)^{-1} \beta_2}\right)^{\alpha_2 - nu}} \right)} \right)^2 \quad (3.2.2.14)$$

Posterior Distribution under Generalized Truncated Geometric and Inverted Gamma Priors

If the prior assumed for shape parameter u is the generalized truncated geometric distribution then the pdf is given by

$$g_{31}(u, \theta_2) = (1 - m + \theta_2) \binom{mu - 1}{u - 1} (1 - \theta_2)^{u-1} \theta_2^{mu-u};$$

$$u = 1, 2, 3, \dots; 0 < \theta_2 < 1. \quad (3.2.3.1)$$

The prior for scale parameter v is assumed to be an inverted Gamma distribution having pdf

$$g_{12}(v; \alpha_1, \beta_1) = \frac{\beta_1^{\alpha_1} v^{-(\alpha_1+1)} \exp(-v^{-1}\beta_1)}{\Gamma(\alpha_1)};$$

$$v > 1, \alpha_1 > 0, \beta_1 > 0 \quad (3.2.3.2)$$

and the joint prior distribution of u and v is defined as:

$$g_3\left(u, \frac{v}{X}\right) = g_{31}(u, \theta_2) \cdot g_{12}(v; \alpha_1, \beta_1). \quad (3.2.3.3)$$

By combining likelihood function (3.1) and joint prior function (8.3), the joint posterior distribution of u and v is (3.2.3.4). The marginal posterior distributions of u and v are (3.2.3.5) and (3.2.3.6). The expression for Bayes estimators u and v under the squared error loss function with their respective posterior variance are (3.2.3.7), (3.2.3.8), (3.2.3.9) and (3.2.3.10).

$$\begin{aligned}
 g_3 \left(u, \frac{v}{X} \right) &= \frac{\binom{\mu u - 1}{u - 1} (1 - \theta_2)^{u-1} \theta_2^{\mu u - u} \exp \left(u \sum_{i=1}^n \ln x_i \right) v^{-(\alpha_1 + nu + 1)} \exp \left\{ -v^{-1} (\beta_1 + \sum_{i=1}^n x_i) \right\}}{(\Gamma(u))^n} \\
 &\quad \sum_{u=1}^{\infty} \left(\frac{\binom{\mu u - 1}{u - 1} (1 - \theta_2)^{u-1} \theta_2^{\mu u - u} \exp \left(u \sum_{i=1}^n \ln x_i \right) \int_0^{\infty} v^{-(\alpha_1 + nu + 1)} \exp \left\{ -v^{-1} (\beta_1 + \sum_{i=1}^n x_i) \right\} db}{(\Gamma(u))^n} \right) \\
 g_3 \left(u, \frac{v}{X} \right) &= \frac{\binom{\mu u - 1}{u - 1} (1 - \theta_2)^{u-1} \theta_2^{\mu u - u} \exp \left(u \sum_{i=1}^n \ln x_i \right) v^{-(\alpha_1 + nu + 1)} \exp \left\{ -v^{-1} (\beta_1 + \sum_{i=1}^n x_i) \right\}}{(\Gamma(u))^n} \\
 &\quad \sum_{u=1}^{\infty} \left(\frac{\binom{\mu u - 1}{u - 1} (1 - \theta_2)^{u-1} \theta_2^{\mu u - u} \exp \left(u \sum_{i=1}^n \ln x_i \right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)
 \end{aligned} \tag{3.2.3.4}$$

$$\begin{aligned}
 g_3 \left(\frac{u}{X} \right) &= \frac{\binom{\mu u - 1}{u - 1} (1 - \theta_2)^{u-1} \theta_2^{\mu u - u} \exp \left(u \sum_{i=1}^n \ln x_i \right) \int_0^{\infty} v^{-(\alpha_1 + nu + 1)} \exp \left\{ -v^{-1} (\beta_1 + \sum_{i=1}^n x_i) \right\} db}{(\Gamma(u))^n} \\
 &\quad \sum_{u=1}^{\infty} \left(\frac{\binom{\mu u - 1}{u - 1} (1 - \theta_2)^{u-1} \theta_2^{\mu u - u} \exp \left(u \sum_{i=1}^n \ln x_i \right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right) \\
 g_3 \left(\frac{u}{X} \right) &= \frac{\binom{\mu u - 1}{u - 1} (1 - \theta_2)^{u-1} \theta_2^{\mu u - u} \exp \left(u \sum_{i=1}^n \ln x_i \right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \\
 &\quad \sum_{u=1}^{\infty} \left(\frac{\binom{\mu u - 1}{u - 1} (1 - \theta_2)^{u-1} \theta_2^{\mu u - u} \exp \left(u \sum_{i=1}^n \ln x_i \right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right); u = 1, 2, 3, \dots
 \end{aligned} \tag{3.2.3.5}$$

$$g_3\left(\frac{v}{X}\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{\binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) v^{-nu}}{(\Gamma(u))^n} \right) \frac{\exp\left\{-v^{-1}(\beta_1 + \sum_{i=1}^n x_i)\right\}}{v^{\alpha_1+1}}}{\sum_{u=1}^{\infty} \left(\frac{\binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1+nu}} \right)}; v > 0 \quad (3.2.3.6)$$

$$\hat{u}_3/X = \frac{\sum_{u=1}^{\infty} \left(\frac{u \binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1+nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1+nu}} \right)} \quad (3.2.3.7)$$

$$\hat{v}_3/X = \frac{\sum_{u=1}^{\infty} \left(\frac{\binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) \int_0^{\infty} v \cdot v^{-(\alpha_1+nu+1)} \exp\left\{-v^{-1}(\beta_1 + \sum_{i=1}^n x_i)\right\} db}{(\Gamma(u))^n} \right)}{\sum_{u=1}^{\infty} \left(\frac{\binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1+nu}} \right)}$$

$$\hat{v}_3/X = \frac{\sum_{u=1}^{\infty} \left(\frac{\binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 1)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1+nu-1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1+nu}} \right)} \quad (3.2.3.8)$$

$$\text{Var}\left(\hat{u}_3/X\right)=\left(\frac{\sum_{u=1}^{\infty}\left(\frac{u^2\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u}\exp\left(u\sum_{i=1}^n\ln x_i\right)\Gamma(\alpha_1+nu)}{(\Gamma(u))^n(\beta_1+\sum_{i=1}^nx_i)^{\alpha_1+nu}}\right)}{\sum_{u=1}^{\infty}\left(\frac{\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u}\exp\left(u\sum_{i=1}^n\ln x_i\right)\Gamma(\alpha_1+nu)}{(\Gamma(u))^n(\beta_1+\sum_{i=1}^nx_i)^{\alpha_1+nu}}\right)}\right)-\left(\frac{\sum_{u=1}^{\infty}\left(\frac{u\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u}\exp\left(u\sum_{i=1}^n\ln x_i\right)\Gamma(\alpha_1+nu)}{(\Gamma(u))^n(\beta_1+\sum_{i=1}^nx_i)^{\alpha_1+nu}}\right)}{\sum_{u=1}^{\infty}\left(\frac{\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u}\exp\left(u\sum_{i=1}^n\ln x_i\right)\Gamma(\alpha_1+nu)}{(\Gamma(u))^n(\beta_1+\sum_{i=1}^nx_i)^{\alpha_1+nu}}\right)}\right)^2\right)\quad (3.2.3.9)$$

$$\text{Var}\left(\hat{v}_3/X\right)=\left(\frac{\sum_{u=1}^{\infty}\left(\frac{\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u}\exp\left(u\sum_{i=1}^n\ln x_i\right)\int_0^{\infty}v^2.v^{-(\alpha_1+nu+1)}\exp\left\{-v^{-1}(\beta_1+\sum_{i=1}^nx_i)\right\}db}{(\Gamma(u))^n}\right)}{\sum_{u=1}^{\infty}\left(\frac{\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u}\exp\left(u\sum_{i=1}^n\ln x_i\right)\Gamma(\alpha_1+nu)}{(\Gamma(u))^n(\beta_1+\sum_{i=1}^nx_i)^{\alpha_1+nu}}\right)}\right)-\left(\frac{\sum_{u=1}^{\infty}\left(\frac{\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u}\exp\left(u\sum_{i=1}^n\ln x_i\right)\Gamma(\alpha_1+nu-1)}{(\Gamma(u))^n(\beta_1+\sum_{i=1}^nx_i)^{\alpha_1+nu-1}}\right)}{\sum_{u=1}^{\infty}\left(\frac{\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u}\exp\left(u\sum_{i=1}^n\ln x_i\right)\Gamma(\alpha_1+nu)}{(\Gamma(u))^n(\beta_1+\sum_{i=1}^nx_i)^{\alpha_1+nu}}\right)}\right)^2\right)\quad (3.2.3.10)$$

$$\text{Var}\left(\hat{v}_3/X\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{\binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 2)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu - 2}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)} - \left[\frac{\sum_{u=1}^{\infty} \left(\frac{\binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 1)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu - 1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{\binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)} \right]^2 \quad (3.2.3.10)$$

Special Cases

For $m = 1$, the generalized truncated geometric distribution reduces to a truncated geometric distribution, therefore, the Bayes estimator for scale parameter u and shape parameter v with their respective posterior variance are given by

$$\hat{u}_3^*/X = \frac{\sum_{u=1}^{\infty} \left(\frac{u (1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)} \quad (3.2.3.11)$$

and

$$\hat{v}_3^*/X = \frac{\sum_{u=1}^{\infty} \left(\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 1)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu - 1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)} \quad (3.2.3.12)$$

and in (3.2.3.13) and (3.2.3.14).

$$\text{Var}\left(\hat{u}_3^*/X\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{u^2 (1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)} - \left[\frac{\sum_{u=1}^{\infty} \left(\frac{u(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)} \right]^2 \quad (3.2.3.13)$$

$$\text{Var}\left(\hat{v}_3^*/X\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 2)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu - 2}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)} - \left[\frac{\sum_{u=1}^{\infty} \left(\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu - 1)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu - 1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) \Gamma(\alpha_1 + nu)}{(\Gamma(u))^n (\beta_1 + \sum_{i=1}^n x_i)^{\alpha_1 + nu}} \right)} \right]^2 \quad (3.2.3.14)$$

Posterior Distribution under Generalized Truncated Geometric and Gamma Priors

If the prior assumed for shape parameter u is the generalized truncated geometric distribution, then the pdf is

$$g_{31}(u, \theta_2) = (1 - m + \theta_2) \binom{mu-1}{u-1} (1 - \theta_2)^{u-1} \theta_2^{mu-u};$$

$$u = 1, 2, 3, \dots; 0 < \theta_2 < 1$$

(3.2.4.1)

and the prior for scale parameter v is assumed to be a Gamma distribution with pdf

$$g_{21}(v; \alpha_2, \beta_2) = \frac{\beta_2^{\alpha_2} v^{\alpha_2-1} \exp(-v\beta_2)}{\Gamma(\alpha_2)};$$

$$v > 1, \alpha_2 > 0, \beta_2 > 0.$$

(3.2.4.2)

The joint prior of u and v is defined as:

$$g_4(u, v) = g_{31}(u, \theta_2) \cdot g_{21}(v; \alpha_2, \beta_2). \quad (3.2.4.3)$$

By combining likelihood function (3.1) with joint prior function (9.3), the joint posterior distribution of u and v is given by (3.2.4.4). The marginal posterior distribution of u and v are given by (3.2.4.5) and (3.2.4.6). The expression for Bayes estimators for u and v with their respective posterior variance are given by (3.2.4.7), (3.2.4.8), (3.2.4.9) and (3.2.4.10).

$$g_4(u, v/X) = \frac{\binom{mu-1}{u-1} (1 - \theta_2)^{u-1} \theta_2^{mu-u} v^{(\alpha_2 - nu) - 1} \exp\left\{-\frac{1}{2} \left((2\beta_2)v + \frac{(2\sum_{i=1}^n x_i)}{v} \right)\right\}}{(\Gamma(u))^n \exp\left(-u \sum_{i=1}^n \ln x_i\right)}$$

$$\sum_{u=1}^{\infty} \left(\frac{\binom{mu-1}{u-1} (1 - \theta_2)^{u-1} \theta_2^{mu-u}}{(\Gamma(u))^n \exp\left(-u \sum_{i=1}^n \ln x_i\right)} \int_0^{\infty} v^{(\alpha_2 - nu) - 1} \exp\left\{-\frac{1}{2} \left((2\beta_2)v + \frac{(2\sum_{i=1}^n x_i)}{v} \right)\right\} db \right)$$

$$g_4(u, v/X) = \frac{\binom{mu-1}{u-1} (1 - \theta_2)^{u-1} \theta_2^{mu-u} v^{(\alpha_2 - nu) - 1} \exp\left\{-\frac{1}{2} \left((2\beta_2)v + \frac{(2\sum_{i=1}^n x_i)}{v} \right)\right\}}{(\Gamma(u))^n \exp\left(-u \sum_{i=1}^n \ln x_i\right)}$$

$$\sum_{u=1}^{\infty} \left(\frac{2 \binom{mu-1}{u-1} (1 - \theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2 - nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)} \right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2 \right)^{\alpha_2 - nu}} \right)$$

(3.2.4.4)

$$g_4(u/X) = \frac{\frac{\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u}}{(\Gamma(u))^n \exp(-u \sum_{i=1}^n \ln x_i)} \int_0^\infty v^{(a_2-nu)-1} \exp\left\{\frac{-1}{2}\left((2\beta_2)v + \frac{(2\sum_{i=1}^n x_i)}{v}\right)\right\} dv}{\sum_{u=1}^\infty \left(\frac{2\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u} \exp(u \sum_{i=1}^n \ln x_i) K_{a_2-nu}\left(2\sqrt{\beta_2(\sum x_i)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{a_2-nu}} \right)}$$

$$g_4(u/X) = \frac{2\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u} \exp(u \sum_{i=1}^n \ln x_i) K_{a_2-nu}\left(2\sqrt{\beta_2(\sum x_i)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{a_2-nu}};$$

$$u = 1, 2, 3, \dots$$

(3.2.4.5)

$$g_4(v/X) = \frac{\sum_{u=1}^\infty \left(\frac{\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u} v^{(a_2-nu)-1} \exp\left\{\frac{-1}{2}\left((2\beta_2)v + \frac{(2\sum_{i=1}^n x_i)}{v}\right)\right\}}{(\Gamma(u))^n \exp(-u \sum_{i=1}^n \ln x_i)} \right)}{\sum_{u=1}^\infty \left(\frac{2\binom{\mu-1}{u-1}(1-\theta_2)^{u-1}\theta_2^{\mu-u} \exp(u \sum_{i=1}^n \ln x_i) K_{a_2-nu}\left(2\sqrt{\beta_2(\sum x_i)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{a_2-nu}} \right)}; v > 0$$

(3.2.4.6)

$$\hat{u}_3/X = \frac{\sum_{u=1}^{\infty} \left(\frac{2u \binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)} \quad (3.2.4.7)$$

$$\hat{v}_3/X = \frac{\sum_{u=1}^{\infty} \left(\frac{\binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u}}{(\Gamma(u))^n \exp\left(-u \sum_{i=1}^n \ln x_i\right)} \int_0^{\infty} v \cdot v^{(\alpha_2-nu)-1} \exp\left\{\frac{-1}{2} \left((2\beta_2) v + \frac{(2 \sum_{i=1}^n x_i)}{v} \right)\right\} db \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)}$$

$$\hat{v}_3/X = \frac{\sum_{u=1}^{\infty} \left(\frac{2 \binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu+1} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu+1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \binom{mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)} \quad (3.2.4.8)$$

$$\text{Var}\left(\hat{\mathbf{u}}_4/X\right) = \left[\frac{\sum_{u=1}^{\infty} \left(\frac{2u^2 \binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)} \right] - \left[\frac{\sum_{u=1}^{\infty} \left(\frac{2u \binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)} \right]^2 \quad (3.2.4.9)$$

$$\text{Var}\left(\hat{\mathbf{v}}_4/X\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{2 \binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu+2} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2-nu+2}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)} - \left[\frac{\sum_{u=1}^{\infty} \left(\frac{2 \binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu+1} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2-nu+1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2 \binom{\mu-1}{u-1} (1-\theta_2)^{u-1} \theta_2^{\mu-u} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)} \right]^2 \quad (3.2.4.10)$$

Special Cases

For $m = 1$, the generalized truncated geometric distribution reduces to truncated geometric distribution, therefore, the Bayes estimator for scale parameter u and shape parameter v with their respective posterior variance is given by (3.2.4.11), (3.2.4.12), (3.2.4.13) and (3.2.4.14).

References

Bhattacharyya, S. K., & Singh, N. K. (1994). Bayesian estimation of the traffic intensity in M/Ek/1 queue. *Far. East, Journal of Math and Science*, 2, 57-62.

Brockmeyer, E., Halstorm, H. L., & Jenson, A. (1948). The life and works of A. K. Erlang. *Transactions of the Danish Academy of Technical Sciences*, 2, 277.

Damodaran, D., Gopal, G., & Kapur, P. K. (2010). A Bayesian Erlang software reliability model. *Communication in Dependability and Quality Management*, 13(4), 82-90.

Erlang, A. K. (1909). *The theory of probabilities and telephone conversations*. Nyt Tidsskrift for Matematik B, 20(6), 87-98.

Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical distributions*, 3rd Ed. New York, NY: John Wiley and Sons, Inc.

Haq, A., & Dey, S. (2001). Bayesian estimation of Erlang distribution under different prior distributions. *Journal of Reliability and Statistical Studies*, 4(1), 1-30.

$$\hat{u}_3^* / X = \frac{\sum_{u=1}^{\infty} \left(\frac{2u(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2 - nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2 - nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2u(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2 - nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2 - nu}} \right)} \quad (3.2.4.11)$$

$$\hat{v}_3^* / X = \frac{\sum_{u=1}^{\infty} \left(\frac{2(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2 - nu + 1} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2 - nu + 1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2 - nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left(\left(\sum x_i\right)^{-1} \beta_2\right)^{\alpha_2 - nu}} \right)} \quad (3.2.4.12)$$

$$\mathbf{Var}\left(\hat{u}_4^*/X\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{2u^2(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2u(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)} - \left[\frac{\sum_{u=1}^{\infty} \left(\frac{2u(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2u(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)} \right]^2 \quad (3.2.4.13)$$

$$\mathbf{Var}\left(\hat{v}_4^*/X\right) = \frac{\sum_{u=1}^{\infty} \left(\frac{2(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu+2} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu+2}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)} - \left[\frac{\sum_{u=1}^{\infty} \left(\frac{2(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu+1} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu+1}} \right)}{\sum_{u=1}^{\infty} \left(\frac{2(1-\theta_2)^{u-1} \exp\left(u \sum_{i=1}^n \ln x_i\right) K_{\alpha_2-nu} \left(2\sqrt{\beta_2 \left(\sum x_i\right)}\right)}{(\Gamma(u))^n \left((\sum x_i)^{-1} \beta_2\right)^{\alpha_2-nu}} \right)} \right]^2 \quad (3.2.4.14)$$

Harischandra, K., & Rao, S. S. (1988). A note on statistical inference about the traffic intensity parameter in M/Ek/1 queue. *Sankhya, B*, 50, 144-148.

Jain, S. (2001). *Estimating the change point of Erlang interarrival time distribution*. 39(2), 200-207: INFOR-OTTAWA- Technology Publications. University of Toronto Press, Canada.

Nair, U. N, Muraleedharan, G., & Kurup, P. G. (2003). Erlang distribution model for ocean wave periods. *Journal of the Indian Geophysical Union*, 7(2), 59-70.

Suri, P. K, Bhushan, B., & Jolly, A. (2009). Time estimation for project management life cycles: A simulation approach. *International Journal of Computer Science and Network Security*, 9(5), 211-215.

Wiper, M. P. (1998). Bayesian analysis of Er/M/1 and Er/M/C queues. *Journal of Statistical Planning and Inference*, 69, 65-79.

Acknowledgements

The authors would like to thanks the referees for their careful reading of the manuscript and for their valuable comments and suggestions.

A Proposed Ridge Parameter to Improve the Least Squares Estimator

Ghadban Khalaf
King Khalid University,
Saudi Arabia

Ridge regression, a form of biased linear estimation, is a more appropriate technique than ordinary least squares (OLS) estimation in the case of highly intercorrelated explanatory variables in the linear regression model $\vec{Y} = X\vec{\beta} + \vec{u}$. Two proposed ridge regression parameters from the mean square error (MSE) perspective are evaluated. A simulation study was conducted to demonstrate the performance of the proposed estimators compared to the OLS, HK and HKB estimators. Results show that the suggested estimators outperform the OLS and the other estimators regarding the ridge parameters in all situations examined.

Key words: Multicollinearity, ridge regression, Monte Carlo simulation.

Introduction

Consider the standard model for multiple linear regression

$$\vec{Y} = \beta_0 \mathbf{1} + X\vec{\beta} + \vec{u}, \quad (1)$$

where $\vec{Y}$ is a $(n \times 1)$ column vector of observations on the dependent variable, β_0 is a scalar intercept, $\mathbf{1}$ is a $(n \times 1)$ vector with all components equal to unity, X is a $(n \times p)$ fixed matrix of observations on the explanatory variables and is of full rank p , $\vec{\beta}$ is a $(p \times 1)$ unknown column vector of regression coefficients and $\vec{u}$ is a $(n \times 1)$ vector of random errors, $E(\vec{u}) = 0$, $E(\vec{u}\vec{u}') = \sigma^2 I_n$, where I_n denotes the $(n \times n)$ identity matrix and the prime denotes the transpose of a matrix.

The OLS estimator, $\vec{\hat{\beta}}$, of the parameters is given by

$$\vec{\hat{\beta}} = (X'X)^{-1} X'\vec{Y} \quad (2)$$

where $\vec{\hat{\beta}}$ is an unbiased estimator of $\vec{\beta}$. Multiple linear regression is very sensitive to predictors that are in a configuration of near collinearity. When this is the case, the model parameters become unstable (large variances) and cannot be interpreted. From a mathematical standpoint, near-collinearity makes the $X'X$ matrix ill-conditioned (with X the data matrix), that is, the value of its determinant is nearly zero, thus, attempts to calculate the inverse of the matrix result in numerical snags with uncertain final values.

Exact collinearity occurs when at least one of the predictors is a linear combination of other predictors. Therefore, X is not a full rank matrix, the determinant of X is exactly zero, and inverting $X'X$ is not simply difficult, it does not exist.

When multicollinearity occurs, the least squares estimates remain unbiased and efficient. The problem is that the estimated standard error of the coefficient β_i (for example, S_{β_i}) tends to be inflated. This standard error has a tendency to be larger than it would be in the absence of multicollinearity because the estimates are very sensitive to any changes in the sample observations or in the model specification. In other words, including or excluding a particular variable or certain observations may greatly

Ghadban Khalaf is an Associate Professor of Statistics in the Department of Mathematics. He is a member of the Faculty of Science. Email him at: albadran50@yahoo.com.

change the estimated partial coefficient. If S_{bi} is larger than it should be, then the t -value for testing the significance of β_i is smaller than it should be. Thus, it becomes more likely to conclude that a variable X_i is not important in a relationship when, in fact, it is important.

Several criteria have been put forth to detect multicollinearity problems. Draper and Smith (1998) suggested the following:

- (1) Check if any regression coefficients have the wrong sign, based on prior knowledge.
- (2) Check if predictors anticipated to be important based on prior knowledge have regression coefficients with small t -statistics.
- (3) Check if deletion of a row or a column of the X matrix produces a large change in the fitted model.
- (4) Check the correlations between all pairs of predictor variables to determine if any are unexpectedly high.
- (5) Examine the variance inflation factor (VIF).
The VIF of X_i is given by:

$$VIF_i = \frac{1}{1 - R_i^2}, \quad (3)$$

where R_i^2 is the squared multiple correlation coefficient resulting from the regression of X_i against all other explanatory variables.

If X_i has a strong linear relation with other explanatory variables, then R_i^2 will be close to one and VIF values will tend to be very high. However, in the absence of any linear relation among explanatory variables, R_i^2 will be zero and the VIF will equal one. It is known that a VIF value greater than one indicates deviation from orthogonality and has tendencies

to col linearity. Leclerc and Pireaux (1995) suggested that a VIF value exceeding 300 may indicate the presence of multicollinearity. Conversely, examining a pairwise correlation matrix of explanatory variables might be insufficient to identify collinearity problems because near linear dependencies may exist among more complex combinations of regressors, that is, pairwise independence does not imply independence. Because VIF is a function of the multiple correlation coefficient among the explanatory variables, it is a much more informative tool for detecting multicollinearity than the simple pairwise correlations.

Many procedures have been suggested in an attempt to overcome the effects of multicollinearity in regression analysis. Horel and Kennard (1970) proposed a class of biased estimator called ridge regression estimators as an alternative to the OLS estimator in the presence of collinearity. Freund and Wilson (1998) summarize these into three classes: variable selection, variable redefinition and biased estimation, such as ridge regression. Ridge regression is a variant of ordinary multiple linear regression whose goal is to circumvent the problem of predictors collinearity. Ridge regression gives up the OLS estimator as a method for estimating the parameters of the model and focuses instead on the $X'X$ matrix; this matrix will be artificially modified in order to make its determinant appreciably different from zero. The idea is to add a small positive quantity, for example k , to each of the diagonal elements of the matrix $X'X$ to reduce linear dependencies observed among its columns. A solution vector is thus obtained by the expression

$$\vec{\beta}^* = (X'X + k I_p)^{-1} X' \vec{Y}, \quad (4)$$

where the ridge parameter $k > 0$ represents the degree of shrinkage. By adding the term $k I_p$, I_p is an identity matrix of the same order as $X'X$, the ridge-regression model reduces multicollinearity and prevents the matrix $X'X$

from being singular even if X itself is not of full rank.

Note that if $k = 0$, the ridge-regression coefficients, defined by (4), are equal to those from the traditional multiple-regression model given by (2). This makes the new model parameters somewhat biased, that is,

$E(\tilde{\beta}^*) \neq \tilde{\beta}$, (whereas the parameters as calculated by the OLS method are unbiased estimators of the true parameters). However, the variances of the new parameters are smaller than that of the OLS parameters and, in fact, so much smaller than their MSE may also be smaller than that of the parameters of the least squares model. This is an illustration of the fact that a biased estimator may outperform an unbiased estimator provided its variance is small enough.

Perhaps the best way for choosing the ridge regression parameter (k) would be to minimize the expected squared difference between the estimate and the parameter being estimated, that is, the MSE. This would reveal the ideal balance between increase in bias and reduction in variance of the estimator, where

$$MSE = Variance + (Bias)^2. \quad (5)$$

Therefore, it is helpful to allow a small bias in order to achieve the main criterion of keeping the MSE small: this is precisely what ridge regression seeks to accomplish.

Several methods for estimating k have been proposed, for example see: Hoerl and Kennard (1970), Hoerl, et al. (1975), McDonald and Galarneau (1975), Lawless and Wang (1976), Hocking, et al. (1976), Wichern and Churchill (1978), Nordberg (1982), Saleh and Kibria (1993), Singh and Tracy (1999), Wencheke (2000), Kibria (2003), Khalaf and Shukur (2005), Alkhamisi, et al. (2006), Alkhamisi & Shukur (2007), Khalaf (2011) and Khalaf, et al., (2012).

The Main Result

Identifying the optimal method for choosing k is beyond the goal of this study; Hoerl and Kennard (1970) showed that the

optimal values for k_i will be

$$\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\beta}_i^2}, \quad (6)$$

$i = 1, 2, \dots, p.$

The acronym HK is used for this estimator. Hoerl and Kennard (1970) stated that “based on experience the best method for achieving a better estimator $\tilde{\beta}^*$ is to use $\hat{k}_i = k$ for all i .”

Thus, the $\hat{k}_i$ – values of (6) can be combined to obtain a single value of k . Thereby it is not advisable to use an ordinary average because a large k and too much bias would result. Hoerl, et al. (1975) proposed a more reasonable averaging, namely the harmonic mean given by

$$\hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}, \quad (7)$$

where p denotes the number of parameters and $\hat{\sigma}^2$ is given by

$$\hat{\sigma}^2 = \frac{RSS}{n - p}, \quad (8)$$

where RSS denotes the residual sum of squares and the acronym HKB is used for estimator (7). The original definition of k provided by Hoerl and Kennard (1970) and Hoerl, et al. (1975) is used throughout this article to suggest the proposed estimators as modifications of their estimators. It is known that the denominator $(n - p + 2)$ yields an estimator of σ^2 with a lower MSE than the unbiased estimator given by (8) (Rao, 1973). Thus, the use of $\hat{\sigma}^2$ is suggested and is defined by

$$\hat{\sigma}^{*2} = \frac{RSS}{n - p + 2}, \quad (9)$$

to estimate $\hat{\sigma}^2$ in both (6) and (7). This leads to the following new estimators

$$\hat{k}_1^* = \frac{\hat{\sigma}^{*2}}{\hat{\beta}_1^2}, \quad (10)$$

$$i = 1, 2, \dots, p$$

and

$$\hat{k}_2^* = \frac{p\hat{\sigma}^{*2}}{\hat{\beta}'\hat{\beta}}. \quad (11)$$

This investigation shows that both $\hat{k}_1^*$ and $\hat{k}_2^*$ in (10) and (11) perform very well relative to the OLS estimator from the MSE point of view.

Methodology

The Simulation

A simulation study was conducted to evaluate the performance of the proposed estimators and to illustrate their superiority. The simulation study concerns a regression model, without the intercept term, with $p = 6$. The simulation procedure suggested by McDonald and Galarnau (1975), Gibbons (1981) and Kibria (2003) was used to generate the explanatory variables:

$$X_{ij} = (1 - \rho^2)^{\frac{1}{2}} z_{ij} + \rho z_{ip}, \quad (12)$$

$$i = 1, 2, \dots, n,$$

$$j = 1, 2, \dots, p,$$

where z_{ij} 's are independent standard normal distribution, ρ^2 is the correlation between any two explanatory variables and p is the number of explanatory variables. The value of ρ^2 is taken as 0.9, 0.99, 0.999 and 0.9999, respectively. The resulting condition numbers (CN) of the generated X equal: 87.36, 368.62, 867.05 and 4250.64, respectively. The n observations for the dependent variable $\bar{Y}$ are determined by:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + u_i, \quad (13)$$

$$i = 1, 2, \dots, n$$

where u_i are independent normal $(0, \sigma^2)$ pseudo-numbers and β_0 is assumed to be identically zero. In this study n is 10, 100 and 1,000 in order to cover both small and large sample sizes. The parameter values were chosen so that $\beta'\beta = 1$, which is a common restriction in simulation studies (Muniz & Kibria, 2009). For given values of p , n and ρ^2 , the experiment was repeated 10,000 times by generating 10,000 samples. For each replicate, the values of k for different proposed estimators and the corresponding ridge estimators were calculated using equation (4) where k is given by (6), (7), (10) and (11).

To investigate whether the ridge estimator is better than the OLS estimator, the MSE was calculated using the equation

$$MSE(\tilde{\beta}^*) = \frac{1}{10000} \sum_{r=1}^{10000} (\beta^* - \beta)'(\beta^* - \beta). \quad (14)$$

Results

Ridge estimators are constructed with the aim of having smaller MSE than the MSE for the least squares. Improvement, if any, can therefore be studied by looking at the amounts of these MSE's. The detailed results of the simulations are shown in Tables 1 – 3. The results concerning the MSE's and the comparisons of ridge estimators with least squares is then dealt with. To summarize these findings:

- (1) Regardless of the condition of $X'X$, the values of MSE of the estimators relative to the OLS estimator are small and therefore the improvement of the ridge estimators over the OLS estimator is remarkable. This may indicate that the influence of multicollinearity upon the MSE criterion is relatively weak. Consequently, the two proposed estimators, given by $\hat{k}_1^*$ and $\hat{k}_2^*$, are far more effective than HK and HKB in improving the OLS estimator.
- (2) Regardless of sample size, the differences of the values of each type of the suggested

estimators are trivial. The $\hat{k}_2^*$ estimator, defined by (11), performed very well; it appears to outperform $\hat{k}_1^*$, and it is also considerably better than HK and HKB.

In summary, the proposed estimators can greatly improve the OLS estimator, as well the HK and HKB estimators, under the MSE criterion. The proposed estimators appear to offer an opportunity for a large reduction in MSE when the degree of multicollinearity as measured by the CN is high.

Table 1: The MSE of the Suggested Estimators, HK, HKB and the OLS Estimator (n = 20)

ρ^2	0.9	0.99	0.999	0.9999
CN	87.36	368.62	867.05	4250.64
OLS	0.190	0.284	0.817	4.213
$\hat{k}_1^*$	0.125	0.156	0.360	1.578
$\hat{k}_2^*$	0.141	0.153	0.240	0.259
HK	0.197	0.207	0.280	0.626
HKB	0.180	0.264	0.688	2.363

Table 2: The MSE of the Suggested Estimators, HK, HKB and the OLS Estimator (n = 100)

ρ^2	0.9	0.99	0.999	0.9999
CN	87.36	368.62	867.05	4250.64
OLS	0.40	0.058	0.169	0.940
$\hat{k}_1^*$	0.034	0.046	0.086	0.360
$\hat{k}_2^*$	0.032	0.036	0.070	0.224
HK	0.045	0.045	0.083	0.250
HKB	0.039	0.056	0.154	0.631

Table 3: The MSE of the Suggested Estimators, HK, HKB and the OLS Estimator (n = 1,000)

ρ^2	0.9	0.99	0.999	0.9999
CN	87.36	368.62	867.05	4250.64
OLS	0.030	0.045	0.130	0.658
$\hat{k}_1^*$	0.026	0.036	0.073	0.229
$\hat{k}_2^*$	0.023	0.028	0.058	0.156
HK	0.027	0.031	0.065	0.183
HKB	0.029	0.044	0.108	0.449

Conclusion

Ridge regression is more than a last resort attempt to salvage least square linear regression in the case of near or full collinearity of predictors. It is to be considered a major linear regression technique that proves its usefulness when collinearity is problematic. From the MSE point of view, it is not surprising that the use of traditional multiple linear regression suffers from multicollinearity problems and clearly shows that ridge regression performs best when the input data are multicollinear.

Two methods for specifying k were proposed herein and were evaluated in terms of MSE via simulation techniques. Comparisons were made with other ridge-type estimators evaluated elsewhere. The simulation study showed that the OLS estimator is dominated by these estimators in all cases investigated and that the improvement of the suggested estimators is substantial from the MSE point of view. Finally, although there are many strategies for choosing an optimal value for k , there is no consensus regarding the best or most general way to choose k . In other words, the best method for estimating k is an unsolved problem and there is no rule for choosing k evaluated to date that assures the corresponding ridge estimator is uniformly better (in terms of MSE) than the OLS estimator.

Acknowledgement

This research was financed by King Khalid University, Abha, Saudi Arabia, Project No. 182.

References

- Alkhamisi, M. A., & Shukur, G. (2007). A Monte Carlo study of recent ridge parameters. *Communications in Statistics - Simulation and Computation*, 36(3), 535-547.
- Alkhamisi, M. A., Khalaf, G., & Shukur, G. (2006). Some modification for choosing ridge parameter. *Communications in Statistics - Theory and Methods*, 35, 2005-2020.
- Draper, N. R., & Smith, H. (1998). *Applied regression analysis*, 3rd Ed. New York, NY: John Wiley and Sons.
- Freund, R. J., & Wilson, W. J. (1998). *Regression analysis: Statistical modeling of a response variable*, 1st Ed. San Diego, CA: Academic Press.
- Gibbons, D. G. (1981). A simulation study of some ridge estimators. *Journal of the American Statistical Association*, 76(373), 131-139.
- Hocking, R. R., Speed, F. M., & Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics*, 18(4), 425-437.
- Horel, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.
- Hoerl, A. E., Kennard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some simulation. *Communications in Statistics*, 4, 105-123.
- Khalaf, G. (2011). Ridge regression: An evaluation to some new modifications. *International Journal of Statistics and Analysis*, 1(4), 325-342.
- Khalaf, G., Månsson, K., Shukur, G., & Sjölander, P. (2012). A Tobit ridge regression estimator. To appear in *Communications in Statistics - Theory and Methods*.
- Khalaf, G., & Shukur, G. (2005). Choosing ridge parameter for regression problems. *Communications in Statistics - Theory and Methods*, 34, 1177-1182.
- Kibria, B. M. (2003). Performance of some new ridge regression estimators. *Communications in Statistics - Simulation and Computation*, 32(2), 419-435.
- Lawless, J. F., & Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics*, A5, 307-323.
- Leclerc, G., & Pireaux, J. J. (1995). The use of least squares for XPS peak parameters estimation, part 3: Multicollinearity, ill-conditioning and constraint-induced bias. *Journal of Electron Spectroscopy Related Phenomena*, 71, 179-190.
- McDonald, G. C., & Galarneau, D. I. (1975). A Monte Carlo Evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70, 407-416.
- Muniz, G., & Kibria, B. M. G. (2009). On some ridge regression estimators: An empirical comparison. *Communications in Statistics - Simulation and Computation*, 38, 621-630.

Nordberg, L. (1982). A Procedure for determination of a good ridge parameter in linear regression. *Communications in Statistics, A11*, 285-309.

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York, NY: John Wiley and Sons.

Saleh, A. K., & Kibria, B. M. (1993). Performances of some new preliminary test ridge regression estimators and their properties. *Communications in Statistics - Theory and Methods*, 22, 2747-2764.

Singh, S., & Tracy, D. S. (1999). Ridge regression using scrambled responses. *Metrika*, b, 147-157.

Wencheko, E. (2000). Estimation of the signal-to-noise in the linear regression model. *Statistical Papers*, 41, 327-343.

Wichern, D., & Churchill, G. (1978). A comparison of ridge estimators. *Technometrics*, 20, 301-311.

Exact Logistic Regression for a Matched Pairs Case-Control Design with Polytomous Exposure Variables

Shyam S. Ganguly
Sultan Qaboos University
Muscat 123, Oman

Logistic regression methods are useful in estimating odds ratios under matched pairs case-control designs when the exposure variable of interest is binary or polytomous in nature. Analysis is typically performed using large sample approximation techniques. When conducting the analysis with polytomous exposure variable, situations where the numbers of discordant pairs in the resulting cells are small or the data structure is sparse can be encountered. In such situations, the asymptotic method of analysis is questionable, thus an exact method of analysis may be more suitable. A method is presented that performs exact inference in the case of pair-wise matched case-control data with more than two unordered exposure categories using a distribution of conditional sufficient statistics of logistic model parameters.

Key words: Conditional logistic regression, sufficient statistic, exact analysis, Diophantine systems.

Introduction

In epidemiological studies, the matched case-control design is often conducted to establish the relationship between disease incidence and an exposure variable of interest in terms of odds ratio (Mantel & Haenszel, 1959; Miettinen, 1970; Ejigou & McHugh, 1977, 1981). The binary logistic model (Cox, 1970) is useful in the estimation of odds ratios under matched pair case-control designs. Prentice (1976), Holford (1978), Holford, et al. (1978), Klinbaum, et al. (1982) and Breslow and Day (1980) provide detailed discussions regarding the estimation of odds ratios using binary logistic models that are conditional on disease status. The polytomous logistic model (Prentice & Pyke, 1979; Dubin & Pasternack, 1988; Liang & Stewart, 1987) has also been found to be useful in estimating odds ratios in the case of matching design when multiple case-control groups are considered; however, when conducting a pair-wise matched

case-control study, a situation where the risk factor under investigation has more than two levels, which may be ordinal or nominal in nature, can be encountered. Ganguly and Naik-Nimbalkar (1995) discuss analysis in the case of a risk factor with a natural ordering, and Ganguly (2006) further estimated the covariate adjusted odds ratios in the case of the ordinal multiple level exposure variables.

Nominal response situations were studied in detail by Pike, et al. (1975), who estimated odds ratios between blood types and development of disease, considering a hypothetical data set. Holford, et al. (1978) analyzed the same data set using a binary logistic model with a conditional likelihood procedure, and Ganguly and Naik-Nimbalkar (1992a, 1992b) further analyzed the data, modeling retrospective probabilities using a polytomous logistic model. All estimation procedures described are based on maximizing the conditional likelihood that relies on asymptotic approximations. The validity of the analysis based on the asymptotic method may be in question when the sample size is small or the data are sparse. In such situations the exact method of analysis is more appropriate (Breslow & Day, 1980; Agresti, 1990; Mehta, 1994).

Shyam S. Ganguly is an Associate Professor of Statistics & Epidemiology in the College of Medicine and Health Sciences in the Department of Family Medicine & Public Health. Email him at: ganguly@squ.edu.om.

Cox (1970) put forth a method for exact logistic regression analysis involving a single parameter in unmatched logistic models, which can also be applied to matched designs when the response is binary. Tritchler (1984) estimated model parameters based on an algorithm developed for a permutation test by Pagano and Tritchler (1983). Hirji, Mehta and Patel (1988) developed a recursive algorithm to compute the exact conditional distribution of sufficient statistics of the parameters involved in a logistic model for analyzing data from a matched case-control design. Hirji (1992) provided an efficient method for computing exact conditional distributions of sufficient statistics for the parameters involved in polytomous response models. However, none of these studies discuss matched case-control designs involving more than two exposure categories. This article proposes a method that uses the conditional distribution of sufficient statistics of logistic model parameters to perform exact inference in the case of a 1-1 matched case-control data with a polytomous exposure variable.

The Logistic Regression Model

Assume k possible levels of an exposure variable of interest. Let p_{ji} be the probability that, in a given pair, the case is exposed to level j and the control is exposed to level i , conditional on one of them being exposed to level j and the other exposed to level i ($1 \leq i < j \leq k$). In addition, let F_1 , be the exposure level associated with the case and F_0 for the control. Consider n_{ij} as the number of case-control pairs in which the case is exposed to level i and the control is exposed to level j and assume that the exposure levels associated with the case and the control are independent. The results of the case-control investigation, in general, may be represented as shown in Table 1.

Let $s_{ij} = n_{ij} + n_{ji}$ ($1 \leq i < j \leq k$) represent the number of discordant pairs referencing the $(i, j)^{\text{th}}$ and $(j, i)^{\text{th}}$ cells in Table 1. Further, consider $Y_{ij\ell}$ as the case-control indicator for the ℓ^{th} pair ($\ell = 1, \dots, s_{ij}$) such that

$$Y_{ij\ell} = \begin{cases} 1 & \text{when case is at } i^{\text{th}} \text{ and control is at } j^{\text{th}} \text{ level} \\ 0 & \text{otherwise} \end{cases}$$

Table 1: Representation of Data from a Matched Pair Study with k Exposure Levels

Exposure Level for Case (F_1)	Exposure Level for Control (F_0)				
	1	2	i	j	k
1	n_{11}	n_{12}	n_{1i}	n_{1j}	n_{1k}
2	n_{21}	n_{22}	n_{2i}	n_{2j}	n_{2k}
i	n_{i1}	n_{i2}	n_{ii}	n_{ij}	n_{ik}
j	n_{j1}	n_{j2}	n_{ji}	n_{jj}	n_{jk}
k	n_{k1}	n_{k2}	n_{ki}	n_{kj}	n_{kk}

Following Ganguly and Naik-Nimbalkar (1992a) the probability that, in a given pair, the case is exposed to level j and the control is exposed to level i conditional on the fact that one is exposed to level j and the other is exposed to level i , is given by

$$p_{ji\ell} = p_{ji} = \frac{\exp(\alpha_j - \alpha_i)}{1 + \exp(\alpha_j - \alpha_i)}, \quad (1)$$

with

$$p_{ji} + p_{ij} = 1, \quad 1 \leq i < j \leq k, \quad \ell = 1, \dots, s_{ij}.$$

The parameter α_j ($j = 1, \dots, k$) describes the additional exposure for an individual in the j^{th} category for becoming a case. The odds ratios, for comparing categories j and i , under model (1) is given by

$$r_{ji} = \exp(\alpha_j - \alpha_i), \quad (1 \leq i < j \leq k).$$

Exact Conditional Distribution of Sufficient Statistics

If the observed discordant case-control pairs in the $(i, j)^{\text{th}}$ and $(j, i)^{\text{th}}$ are considered as

$y_{ij1}, y_{ij2}, \dots, y_{ij\ell}, \dots, y_{ijs_{ij}}$ and assumed to be independent, then the likelihood L_{ij} for the s_{ij} pairs is conditional on the study design and is given by

$$L_{ij} = \text{pr} \left[Y_{ij1} = y_{ij1}, \dots, Y_{ij\ell} = y_{ij\ell}, \dots, Y_{ijs_{ij}} = y_{ijs_{ij}} \mid S_{ij} = s_{ij} \right] \\ = \prod_{\ell=1}^{s_{ij}} (p_{ij\ell})^{y_{ij\ell}} (1 - p_{ij\ell})^{1-y_{ij\ell}} \quad (2)$$

Using relation (1) and (2), L_{ij} is given by

$$L_{ij} = \frac{\exp \left[n_{ij}(\alpha_i - \alpha_j) \right]}{\prod_{\ell=1}^{s_{ij}} [1 + \exp(\alpha_i - \alpha_j)]}, \quad (3)$$

therefore, the overall conditional likelihood L is given by

$$L = \prod_{i=1}^{k-1} \prod_{j=i+1}^k L_{ij} \\ = \frac{\exp \left[\sum_{i=1}^{k-1} \alpha_i t_i \right]}{\prod_{i=1}^{k-1} \prod_{j=i+1}^k [1 + \exp(\alpha_i - \alpha_j)]^{s_{ij}}}, \quad (4)$$

where

$$\sum_{i=1}^k \alpha_i = 0$$

or

$$\alpha_k = -\sum_{i=1}^{k-1} \alpha_i$$

and

$$t_i = \sum_{j=i+1}^k n_{ij} - \sum_{j=1}^{i-1} n_{ji} + \sum_{j=1}^{i-1} n_{jk}, \\ i = 1, \dots, k-1.$$

The likelihood (4) can also be represented as

$$L = H(\alpha_1, \dots, \alpha_{k-1}) \left\{ \exp \left(\sum_{i=1}^{k-1} \alpha_i t_i \right) \right\}, \quad (5)$$

where

$$H(\alpha_1, \dots, \alpha_{k-1}) = \frac{1}{\prod_{i < j} \prod [1 + \exp(\alpha_i - \alpha_j)]^{s_{ij}}}$$

Relation (5) shows an exponential family of dimension $k-1$ and the T_i 's are jointly sufficient for α_i ($i = 1, \dots, k-1$), whose joint distribution is obtained by summing over all n_{ij} values, such that $T_i = t_i$ ($i = 1, \dots, k-1$) and $S_{ij} = s_{ij}$. The joint distribution is thus given by

$$\text{pr} (T_1 = t_1, \dots, T_{k-1} = t_{k-1} \mid S_{ij} = s_{ij}, i < j = 1, \dots, k) \\ = \frac{C(t_1, \dots, t_{k-1}) \exp \left[\sum_{i=1}^{k-1} \alpha_i t_i \right]}{\prod_{i=1}^{k-1} \prod_{j=i+1}^k [1 + \exp(\alpha_i - \alpha_j)]^{s_{ij}}}, \quad (6)$$

where $C(t_1, \dots, t_{k-1})$ is the number of distinct set of values assumed by n_{ij} which yield the values $t_1, \dots, t_{k-1}$ for the joint sufficient statistic.

Following Cox (1970), the natural statistic for making an inference about α_{k-1} , for example, in the presence of $\alpha_1, \dots, \alpha_{k-2}$, is T_{k-1} , conditioned on $T_{k-2}, \dots, T_1$ and $S_{ij} = n_{ij} + n_{ji}$. This conditional distribution is given by

$$p_r (T_{k-1} = t_{k-1} \mid T_1 = t_1, \dots, T_{k-2} = t_{k-2}, S_{ij} = s_{ij}, i < j = 1, \dots, k) \\ = \frac{p_r (T_1 = t_1, \dots, T_{k-2} = t_{k-2}, T_{k-1} = t_{k-1} \mid S_{ij} = s_{ij}, i < j = 1, \dots, k)}{p_r (T_1 = t_1, \dots, T_{k-2} = t_{k-2}, S_{ij} = s_{ij}, i < j = 1, \dots, k)} \quad (7)$$

The distribution in the denominator of (7) is obtained by summing (6) over all possible t_{k-1} and is given by

$$\begin{aligned}
 p_r(T_1 = t_1, \dots, T_{k-2} = t_{k-2} | S_{ij} = s_{ij}, i < j = 1, \dots, k) \\
 = \sum_u \frac{C(t_1, \dots, t_{k-2}, u) \exp[\sum_{i=1}^{k-1} \alpha_i t_i]}{\prod_{i=1}^{k-1} \prod_{j=i+1}^k [1 + \exp(\alpha_i - \alpha_j)]^{s_{ij}}}, \\
 = \frac{\sum_u C(t_1, \dots, t_{k-2}, u) \exp[\sum_{i=1}^{k-1} \alpha_i t_i]}{\prod_{i=1}^{k-1} \prod_{j=i+1}^k [1 + \exp(\alpha_i - \alpha_j)]^{s_{ij}}}.
 \end{aligned} \quad (8)$$

From (6) and (8) the conditional distribution (7) is obtained and is given by

$$\begin{aligned}
 \Pr(T_{k-1} = t_{k-1} | T_1 = t_1, \dots, T_{k-2} = t_{k-2}, S_{ij} = s_{ij}, i < j = 1, \dots, k) \\
 = \frac{C(t_1, \dots, t_{k-1}) \exp \sum_{i=1}^{k-1} \alpha_i t_i}{\sum_u C(t_1, \dots, t_{k-2}, u) \exp[\sum_{i=1}^{k-2} \alpha_i t_i + \alpha_{k-1} u]}, \\
 = \frac{C(t_1, \dots, t_{k-1}) \exp(\alpha_{k-1} t_{k-1})}{\sum_u C(t_1, \dots, t_{k-2}, u) \exp(\alpha_{k-1} u)}
 \end{aligned} \quad (9)$$

where u is an index ranging over the values taken by T_{k-1} and $C(t_1, \dots, t_{k-2}, u)$ is the number of distinct set of values of n_{ij} ($i < j = 1, \dots, k$) which when substituted in (9) yield $T_1 = t_1, \dots, T_{k-2} = t_{k-2}, T_{k-1} = u$ and $S_{ij} = s_{ij}$. Note that (9) does not involve $\alpha_1, \dots, \alpha_{k-2}$. In order to simplify the notation, denote $(t_1, \dots, t_{k-2})$ by $\underline{t}_{k-2}$, thus the distribution (9) can be written as

$$p_r(t_{k-1}; \alpha_{k-1}) = \frac{C(\underline{t}_{k-2}, t_{k-1}) \exp(\alpha_{k-1} t_{k-1})}{\sum_u C(\underline{t}_{k-2}, u) \exp(\alpha_{k-1} u)}. \quad (10)$$

An important case of (10) corresponds to $\alpha_{k-1} = 0$,

$$p_r(t_{k-1}; 0) = \frac{C(\underline{t}_{k-2}, t_{k-1})}{\sum_u C(\underline{t}_{k-2}, u)} \quad (11)$$

so that the distribution is determined by the combinatorial coefficients. The computation of the combinatorial coefficient $C(\underline{t}_{k-2}, t_{k-1})$ involves calculations which are computationally prohibitive for larger value of k , the number of levels of exposure, and with small numbers of discordant pairs in the resulting cells.

The Computational Method

A computational method that can be used for obtaining the combinatorial coefficients involved in the distribution (10) is available. Here, interest lies in computing the coefficients $C(\underline{t}_{k-2}, \cdot)$, where the dot indicates that the corresponding argument varies over its permissible range of values for t_{k-1} . The coefficient $C(\underline{t}_{k-2}, t_{k-1})$ may be counted following the procedure involved in investigating the solutions of the Diophantine systems in non-negative integers as described in Constantine (1987). The Diophantine system is represented by

$$\sum_{r=1}^n a_{ir} x_r = t_i, i = 1, \dots, k-1, \quad (12)$$

where a_{ir} and t_i are non-negative integers. Writing

$$\begin{aligned}
 \underline{x} &= (x_1, \dots, x_n), \\
 \underline{t} &= (t_1, \dots, t_{k-1}), \\
 \underline{\xi} &= (\xi_1, \dots, \xi_{k-1})
 \end{aligned}$$

and

$$\underline{\xi}^t = \prod_{i=1}^{k-1} \xi_i^{t_i}.$$

If $C(\underline{t})$ is the number of solutions to (12), then using the generating function results in

$$\sum_{t \geq 0} C(t) \underline{\xi}^t = \prod_{r=1}^n (1 - \xi_1^{a_{1r}}, \dots, \xi_{k-1}^{a_{(k-1)r}})^{-1}$$

which is

$$\left[\sum_{t \geq 0} C(t) \underline{\xi}^t \right] \left[\prod_{r=1}^n (1 - \xi_1^{a_{1r}}, \dots, \xi_{k-1}^{a_{(k-1)r}}) \right] = 1, \quad (13)$$

Equating the coefficients of $\underline{\xi}^t$ on both sides results in $C(t)$, which is the value of the combinatorial coefficient $C(t_{k-2}, t_{k-1})$. Note that one of the considerations for computing the coefficients using Diophantine systems is that the a_{ir} 's and t_i are non-negative integers valued with non-zero entry in each column of (a_{ir}) . If necessary, this may be achieved by linear transformation with no effect on inference. The non-negativity of the entities involved insures an almost a finite number of solutions to system (12).

The Case of Three Level Exposure

In the simplest situation the polytomous outcome may be considered with three exposure levels. In this case $k = 3$ with two sufficient statistics from (5) which are:

$$t_1 = n_{12} + 2n_{13} + n_{23} \quad (14)$$

and

$$t_2 = -n_{12} + n_{13} + 2n_{23}. \quad (15)$$

If it is of interest to obtain the distribution of the sufficient statistic T_2 given the observed value of $T_1 = t_1$, then the relation (10) reduces to

$$\Pr(t_2; \alpha_2) = \frac{C(t_1, t_2) \exp(\alpha_2 t_2)}{\sum_u C(t_1, u) \exp(\alpha_2 u)}, \quad (16)$$

where u is an index ranging over all possible values taken by T_2 for given $T_1 = t_1$ and $C(t_1, u)$ is the number of distinct set of values for n_{12} , n_{13} and n_{23} which, when substituted in (14) and (15),

yield $T_1 = t_1$ and $T_2 = u$. The range of u is determined by considering the maximum and minimum values of T_2 , namely the maximum value of $t_2 = S_{13} + 2s_{23}$; the minimum value of t_2 may be considered to be zero.

The combinatorial coefficients involved in (16) are computed using the method for solutions of the Diophantine system of equations (12). In (15), to insure that t_2 is a positive integer, a linear transformation namely, $t_2^* = 2t_1 + t_2$, is considered and provides the Diophantine system of equations:

$$t_1 = x_1 + 2x_2 + x_3 \quad (17)$$

and

$$t_2^* = x_1 + 5x_2 + 4x_3 \quad (18)$$

where $x_1 = n_{12}$, $x_2 = n_{13}$ and $x_3 = n_{23}$ respectively. The (a_{ir}) matrix

$$(a_{ir}) = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 5 & 4 \end{pmatrix}$$

results from relations (17) and (18).

Writing $x = (x_1, x_2, x_3)$, $\underline{t}^* = (t_1, t_2^*)$, $\underline{\xi} = (\xi_1, \xi_2)$ and inserting the values of (a_{ir}) in (13), the generating function reduces to

$$\left[\sum_{t^* \geq 0} C(t_1, t_2^*) \xi_1^{t_1} \xi_2^{t_2^*} \right] \left[\begin{matrix} 1 - \xi_1 \xi_2 + \xi_1^3 \xi_2^6 \\ -\xi_1 \xi_2^4 + \xi_1^3 \xi_2^9 \\ -\xi_1^4 \xi_2^{10} \end{matrix} \right] = 1. \quad (19)$$

The combinatorial coefficient $C(t_1, t_2^*)$ is obtained by equating the coefficients of $\xi_1^{t_1} \xi_2^{t_2^*}$ on both sides of (19). This provided the recurrence relation

$$\begin{aligned} & C(t_1, t_2^*) - C(t_1 - 1, t_2^* - 1) \\ & + C(t_1 - 3, t_2^* - 6) - C(t_1 - 1, t_2^* - 4) \\ & + C(t_1 - 3, t_2^* - 9) - C(t_1 - 4, t_2^* - 10) = 0 \end{aligned} \quad (20)$$

The method is repeated for obtaining the values of $c(t_1, u)$, where $0 \leq u \leq \max(t_2^*)$.

Testing and Estimation

Consider the problem of testing the null hypothesis $H_0: \alpha_{k-1} = \alpha_{k-1}^0$ against the one-sided alternative $H_+: \alpha_{k-1} > \alpha_{k-1}^0$. If $\underline{t} = (t_1, \dots, t_{k-1})^T$ is the observed vector of sufficient statistics, then following Lehmann (1959) the p-value for the uniformly most powerful unbiased test of H_0 against H_+ is obtained using relation (10) and is given by

$$\begin{aligned} p_+(t_{k-1}; \alpha_{k-1}^0) &= p_r(T_{k-1} \geq t_{k-1} | T_1 = t_1, \dots, T_{k-2} = t_{k-2}, S_{ij} = s_{ij}, i < j = 1, \dots, k) \\ &= p_r(T_{k-1} \geq t_{k-1} | T_{k-2} = t_{k-2}, S_{ij} = s_{ij}, i < j = 1, \dots, k) \\ &= \sum_{v \geq t_{k-1}} \frac{C(t_{k-2}, v) \exp(\alpha_{k-1}^0 v)}{\sum_u C(t_{k-2}, u) \exp(\alpha_{k-1}^0 u)} \\ &= \frac{\sum_{v \geq t_{k-1}} C(t_{k-2}, v) \exp(\alpha_{k-1}^0 v)}{\sum_u C(t_{k-2}, u) \exp(\alpha_{k-1}^0 u)} \end{aligned} \quad (21)$$

Similarly, the p-value for the test of H_0 versus $H_-: \alpha_{k-1} < \alpha_{k-1}^0$ is given by

$$\begin{aligned} p_-(t_{k-1}; \alpha_{k-1}^0) &= \text{pr}(T_{k-1} \leq t_{k-1} | T_{k-2} = t_{k-2}, S_{ij} = s_{ij}, i < j = 1, \dots, k) \\ &= \frac{\sum_{v \leq t_{k-1}} C(t_{k-2}, v) \exp(\alpha_{k-1}^0 v)}{\sum_u C(t_{k-2}, u) \exp(\alpha_{k-1}^0 u)} \end{aligned} \quad (22)$$

According to Cox (1970), the upper $(1-\alpha)$ level confidence limit for α_{k-1} corresponding to the observed value t_{k-1} is the solution α_{k-1}^U to $P_-(t_{k-1}; \alpha_{k-1}^U) = \alpha$ in (22). Similarly, the solution α_{k-1}^L to $P_+(t_{k-1}; \alpha_{k-1}^L) = \alpha$ in (21) gives the lower $(1-\alpha)$ level confidence limit for α_{k-1} . These values are evaluated by solving the equations

numerically. The two sided alternatives may be obtained by

$$p_r(t_{k-1}; \alpha_{k-1}^0) = 2 \text{ Min} \left[P_+(t_{k-1}; \alpha_{k-1}^0), P_-(t_{k-1}; \alpha_{k-1}^0) \right].$$

Tritchler (1984) suggested that the point estimation of the parameter α_{k-1} denoted by $\hat{\alpha}_{k-1}$ is the value which nearly satisfies $P_-(t_{k-1}; \hat{\alpha}_{k-1}) = P_+(t_{k-1}; \hat{\alpha}_{k-1}) = 0.5$. Following similar techniques, the point and interval estimation of the other model parameters can be obtained.

Numerical Example

The methodology of exact analysis described herein is most suitably performed on a computer, however, for illustration purposes, consider a hypothetical matched pairs data set involving a response variable taking three levels as shown in Table 2. The estimates and 95 percent confidence limits of the parameters for the data set shown in Table 2 were obtained using the exact method of analysis; results are presented in Table 3.

Table 2: Frequency Distribution of Case-Control Pairs with Three Exposure Levels

Exposure Level for Case (F_1)	Exposure Level for Control (F_0)		
	1	2	3
1	20	3	2
2	2	15	2
3	3	2	10

Table 3: Results of Logistic Analysis of the Data in Table 2 Based on Exact Method

Parameters	Exact Estimate	95 Percent Confidence Limits
α_1	-0.120	(-0.630, 0.270)
α_2	-0.150	(-1.050, 0.350)
α_3	0.270	(-0.620, 1.680)

The data set used in the example was considerably large with 59 matched case-control pairs; however, the number of discordant pairs of observations involved in the analysis was very small. Hence, in this situation, the exact method of analysis may be more appropriate. This article considered an exact method of analysis in case of 1-1 matched case-control data when the risk factor of interest is polytomous in nature.

References

- Agresti, A. (1990). *Categorical data analysis*. New York, NY: J. Wiley and Sons.
- Breslow, N. E., & Day N. E. (1980). *Statistical methods in cancer research, Vol. 1: The analysis of case and control studies*. Lyon: International Agency for Research on Cancer.
- Cox, D. R. (1970). *Analysis of binary data*. Mathuen: London.
- Constantine, G. M. (1987). *Combinatorial theory and statistical design*. New York, NY: J. Wiley and Sons.
- Dubin, N., & Pasternack, B. S. (1986). Risk assessment for case-control subgroups by polytomous logistic regression. *American Journal of Epidemiology*, 123, 1101-1117.
- Ejigou, A., & McHugh, R. (1977). Estimation of relative risk from matched pairs in epidemiologic research. *Biometrics*, 33, 552-558.
- Ejigou, A., & McHugh, R. (1981). Relative risk estimation under multiple matching. *Biometrika*, 68, 85-91.
- Ganguly, S. S., & Naik-Nimbalkar, U. (1992). Use of polytomous logistic model in matched case-control studies. *Biometrical Journal*, 34, 209-217.
- Ganguly, S. S., & Naik-Nimbalkar, U. (1992). Polytomous logistic model for matched case-control design. *Journal of Applied Statistics*, 19(4), 455-464.
- Ganguly, S. S., & Naik-Nimbalkar, U. V. (1995). Analysis of ordinal data in a study of endometrial cancer under a matched pairs case-control design. *Statistics in Medicine*, 14, 1545-1552.
- Ganguly S. S. (2006). Cumulative logit models for matched pairs case-control design: Studies with covariates. *Journal of Applied Statistics*, 33(5), 513-522.
- Hirji, K. F. (1992). Computing exact distributions for polytomous response data. *Journal of the American Statistical Association*, 87, 487-492.
- Hirji, K. F, Mehta, C. R., & Patel, N. R. (1988). Exact inference for matched case-control studies. *Biometrics*, 44, 803-814.
- Holford, T. R. (1978). The analysis of pair-matched case-control studies, a multivariate approach, *Biometrics*, 34, 665-672.
- Holford, T. R., While, C., & Kelsey, J. L. (1978). Multivariate analysis for matched case-control studies. *American Journal of Epidemiology*, 107, 246-256.
- Kleinbaum, D. G., Kupper, L. L., & Chambless, L. E. (1982). Logistic regression analysis of epidemiologic data: Theory and practice. *Communications in Statistics - Theory and Methods*, 11(5), 485-547.
- Lehmann, E. L. (1959). *Testing statistical hypothesis*. New York, NY: J. Wiley and Sons.
- Liang, K. Y., & Stewart, W. F. (1987). Polytomous logistic regression for matched case-control studies with multiple case or control groups. *American Journal of Epidemiology*, 125, 720-730.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute*, 22, 719-749.
- Mehta, C. R. (1994). The exact analysis of contingency tables in medical research. *Statistical Methods in Medical Research*, 3, 135-156.
- Miettinen, O. (1970). Estimation of relative risk from individually matched series. *Biometrics*, 26, 75-86.
- Pagano, M., & Tritchler, D. (1983). On obtaining permutation distributions in polynomial time. *Journal of the American Statistical Association*, 78, 435-440.
- Pike, M. C., Casagrande, J., & Smith, P. G. (1975). Statistical analysis of individually matched case-control studies in epidemiology: factor under study a discrete variable taking multiple values. *British Journal of Preventive and Social Medicine*, 29, 196-201.
- Prentice, R. (1976). Use of the logistic model in retrospective studies. *Biometrics*, 32, 599-606.
- Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrics*, 66, 403-411.
- Tritetler, D (1984). An algorithm for exact logistic regression. *Journal of the American Statistical Association*, 79, 709-711.

Graphical Modeling for High Dimensional Data

Munni Begum Jay Bagga C. Ann Blakey
Ball State University,
Muncie, IN

With advances in science and information technologies, many scientific fields are able to meet the challenges of managing and analyzing high-dimensional data. A so-called large p small n problem arises when the number of experimental units, n , is equal to or smaller than the number of features, p . A methodology based on probability and graph theory, termed graphical models, is applied to study the structure and inference of such high-dimensional data.

Key words: High dimensional data, graphical Markov models, conditional independence, Markov properties, chain graphs.

Introduction

Graphical models are the result of a marriage between probability distribution theory and graph theory; these models have been used to study the associations among stochastic variables for decades in many disciplines. Graphical model methodologies evolved through a blend of statistical techniques: log-linear and covariance selection models with constructs of path analysis and the concept of conditional independence (Whitaker, 1990; Edwards, 2000).

Classical examples of graphical model applications include: fitting complex patterns of associations among the factors cross-classifying multidimensional contingency tables, and studying relationships among variables using their covariance structure. The current state of the science of this area includes general methodology on the structural properties of graphical models suggested by the conditional independence and Markov properties.

Conditional independence and Markov properties of graphical models are keys to developing methodologies for high dimensional data analysis in a growing number of computational science fields. Structural learning and computational techniques/algorithms with running time and space complexity are of significant interest. This article outlines a method to address the challenge of making efficient statistical inferences with high dimensional data using the elegant features of graphical models.

Munni Begum is an Associate Professor of Statistics in the Department of Mathematical Sciences. Her research interests include biostatistics methods and applications, computational statistics, graphical models, Bayesian inference and longitudinal and survival data analysis. Email her at: mbegum@bsu.edu. Jay Bagga is a Professor of Computer Science at Ball State University. His research interests include bioinformatics, bioinformatics algorithms, graphical models, graph theory and graph algorithms. Email him at: jbagga@bsu.edu. C. Ann Blakey is an Associate Professor of Genetics in the Department of Biology. Her research interests include molecular genetics, genomics and bioinformatics, biocomputational population genetics modeling and applications. Email her at: ablakey@bsu.edu.

Dimension Reduction Using Regression and Classification

Advances in science and information technology have allowed many scientific fields, such as bioinformatics, computational biology, medicine, pharmacology and toxicology to produce high-dimensional data at an astounding rate. The common scenario of a small number of features p from a large number of experimental units n has resulted in what is termed a large p

small n problem, as the number of experimental units n is equal to or even smaller than the number of features (or parameters) p . To address this problem there has been a surge in research activities offering data reduction methods and statistical inference.

Modern methods for subset selection include the least absolute shrinkage and selection operator, or LASSO (Tibshirani, 1996) method, which provides a sparse solution by picking influential regression coefficients and shrinking others to zero. The LASSO method reduces the dimension from p to $k < p$, k being a subset of the features. A major concern with this method – as well as other related methods such as Garotte (Breiman, 1993) and ridge regression – is whether they successfully identify the correct subset of non-zero regression coefficients. Thus, the question remains as to whether it is possible to achieve a proper projection of the coefficient matrix onto a computationally feasible lower dimension. Even with a substantial dimension reduction, questions still remain as to whether sufficient reduction was achieved.

The idea of sufficient dimension reduction in regression is addressed along the similar line of Fisher's sufficient statistics. Inverse regression (Cook, 2007) based on the principle component regression model has been applied to achieve a sufficiently reduced subspace of predictor variables. If X is a predictor vector in $\mathcal{R}^p$ and Y is the response variable, then a sufficient reduction $R: \mathcal{R}^p \rightarrow \mathcal{R}^q$, $q \leq p$ implies that at least one of the statements (1) $X|Y, R(X) \sim X|R(X)$, (2) $Y|X \sim Y|R(X)$ or (3) $Y \perp\!\!\!\perp X|R(X)$ holds (Cook, 2007). Whereas statements (i) and (ii) correspond directly to inverse and forward regression respectively, statement (iii) connotes the conditional independence of Y and X given the reduced predictor subspace, which is the basis for graphical models. This also implies no loss of relevant information under the sufficient reduction of predictor space. Sparse additive models (Ravikumar, et al., 2007) based on the so-called generalized additive models (Hastie & Tibshirani, 1990) and Bayesian additive regression trees (Chipman, et al., 2009) are other modern data reduction methods in linear and generalized linear regression problem settings.

Dimension Reduction using Graphical Models

Graphical models can be used as efficient tools to investigate the dependence structure of a large number of attributes. Probabilistic expert systems and Bayesian networks (Neapolitan, 1990, 2004, 2009; Cowell et al., 1999) based on directed acyclic graphs are commonly used graphical models in medical diagnosis, disease spread modeling and gene interaction and protein interaction networks. Advancements in the mathematical theory of general graphical models over the past decade through the pioneering work of Lauritzen (1996), Wermuth and Lauritzen (1983), Frydenberg (1990), Andersson (1993, 1995, 1997), Madigan (1995, 1997) and Perlman (1993, 1995, 1997) among others, facilitates the development of a general methodology for practical problems in diverse scientific fields.

Graphical models are a flexible class of models based on both graph and probability theory and can capture complex dependence structure among a large number of stochastic variables efficiently. Although the general methodology is well developed, only a handful of these methods are implemented in practice. There is a need for addressing the computational aspects of these models under a general framework.

This research is based on the challenges of analyzing a large volume of high dimensional molecular interaction data. For example, the interaction between genes, proteins and metabolites are the focus of such emerging fields as transcriptomics, proteomics and metabolomics. A complete biological network consists of all of these interacting components. To understand this complex system it is necessary to apply the divide and conquer rule: break the system into small parts and map out the interactions. Each of these smaller components may be regarded as a unique complex network; thus gene, protein and metabolic interaction networks can be studied under a single framework. Such an endeavor will help address the challenges of high dimensional data analysis and statistical inference.

Preliminaries on Graphical Models

The fields of graph theory and probability theory are well developed. Graph theory is generally studied as a branch of discrete and combinatorial mathematics, however, graph theory and graph algorithms provide an applicable framework in many fields including computer science, mathematical and statistical sciences, biological and chemical sciences and several branches of engineering. The notion of graph has been tied with conditional independence among stochastic variables and their Markov properties. In graphical modeling and localized computations for probabilistic inference, Markov properties play a fundamental role. General chain graphs and their specializations, directed and undirected graphs, each have different types of Markov properties and conditional independence is a common theoretical tool to investigate these fundamental properties of a class of graphs.

In its simplest form, a graph $G = (V, E)$ constitutes a finite set of vertices $V = \{1, 2, \dots, v\}$ and a set of edges $E \subseteq V \times V$. Each edge is thus a pair of vertices $(u, v) \in E$ that incorporates a relationship between two vertices. In a graphical model, the vertices may represent discrete or continuous variables and the edges, which may be undirected or directed, represent conditional dependence. A directed graph $G_D = (V, E)$ contains only directed edges drawn as arrows, where V is a set of vertices and E is a set of ordered pair vertices. Directed graphs, with no directed cycles, are known as directed acyclic

graphs (DAGs) and play a significant role in causal inference. In an undirected graph $G_U = (V, E)$, the edges are undirected and are used mainly to study the association among attributes. Chain graphs (Lauritzen, 1996) have both directed and undirected edges. For a chain graph $G_{ch} = (V, E)$, the vertex set V can be partitioned into numbered subsets that form a dependence chain $V = V_1 \cup V_2 \cup \dots \cup V_T$, such that all the edges between vertices in the same subset are undirected and all edges between different subsets are directed, pointing from a set with lower number to the one with higher number. Figure 1 illustrates directed, undirected and chain graphs (Lauritzen, 1996).

The structural properties of general chain graph models and specialized undirected and directed acyclic graph models are of great interest. Theoretical tools to study the structural properties of a class of graphs are their corresponding Markov properties and characterization of their Markov equivalent classes. An important fact of conditional independence properties in localized computations is that these enable factorization of the joint probability distribution of the random variables associated with the nodes of a graph. Following the notations of Cowell, et al. (1999), let $X_v, v \in V$ be a collection of random variables, taking values in probability spaces $X_v, v \in V$, and let $\mathbf{B}$ be a collection of subsets of V . For $B \in \mathbf{B}$, if $a_B(x)$ denotes a non-negative function that depends only on $x_B = (x_v)_{v \in B}$, then a joint

Figure1: Graphical Model Illustrations

distribution P for X is **B**- hierarchical if its probability density p factorizes as

$$p(x) = \prod_{B \in \mathbf{B}} a_B(x). \quad (1)$$

(Cowell, et al., 1999).

This factorization holds only when $\mathbf{B}$ is a complete subset of the underlying graph. For an undirected graph $G_U = (V, E)$, and a collection of random variables $X_v, v \in V$ taking values in probability spaces $X_v, v \in V$, the joint probability density $p(x)$ is **C** - hierarchical where $\mathbf{C}$ is the set of cliques of G_U . In this case $p(x)$ factorizes as,

$$p(x) = \prod_{C \in \mathbf{C}} \psi_C(x_C), \quad (2)$$

where the function ψ_C is referred to as factor potential of the probability measure P on $\mathbf{C}_v$.

A probability distribution P is said to admit a recursive factorization according to a directed acyclic graph $G_D = (V, E)$ if the joint density $p(x)$ factorizes as,

$$p(x) = \prod_{v \in V} p(x_v | x_{pa(v)}), \quad (3)$$

where $pa(v)$ is the set of parents of the vertex v . If $(u, v) \in E$, but $(v, u) \notin E$, then u is a parent of v and the set of all parents of v is denoted by $pa(v)$. Recursive factorization according to G_D implies **C** hierarchical factorization according to the corresponding undirected moral graph of G_D , denoted as G_D^m . The moralization process of a directed acyclic graph involves adding undirected edges between all pairs of parents of each vertex which are not already joined and then making all edges undirected (Lauritzen, 1996). For a chain graph $G_{ch} = (V, E)$, with dependence chain $V = V_1 \cap V_2 \cap \dots \cap V_T$, the joint density $p(x)$ factorizes as

$$p(x) = \prod_{t=1}^T p(x_{V_t} | x_{C'_{t-1}}), \quad (4)$$

where C'_t are the concurrent variables defined as $C'_t = V_1 \cap V_2 \cap \dots \cap V_t$. If $B'_t = pa(V_t) = bd(V_t)$, then the above factorization reduces to

$$p(x) = \prod_{t=1}^T p(x_{V_t} | x_{B'_t}). \quad (5)$$

For an undirected graph the parent set of a vertex v becomes the neighbor set $nb(v)$. For a chain graph, $bd(v)$ is the set of parents and neighbors of the vertex v . This factorization takes an identical form to that of a directed acyclic graph due to the fact that a chain graph forms a directed acyclic graph of its chain components. One drawback of this representation is that the factorization does not reveal all conditional relationships. To investigate the relationships that are not revealed, if an undirected graph G_{ch}^* with vertex set $V_t \cap B'_t$ is considered, then, for a chain graph, the joint density of a collection of discrete random variables X_v factorizes as,

$$p(x) = \prod_{t=1}^T \frac{p(x_{V_t \cup B'_t})}{p(x_{B'_t})}, \quad (6)$$

and each of the numerators factorizes on the graph G_{ch}^* (Lauritzen, 1996). In addition, if a density $p(x)$ factorizes as in (6), it also factorizes according to the moral graph G_{ch}^m (Lauritzen, 1996).

Associated with a graph G , there are primarily three Markov properties: pairwise, local and global. A probability measure P on X is said to follow the pairwise Markov property relative to G if, for any pair (u, v) of non-adjacent vertices, $u \perp\!\!\!\perp v | V/\{u, v\}$. It follows the local Markov property relative to G , if for any vertex $v \in V$, $v \perp\!\!\!\perp V/cl(v) | bd(v)$. Here the closure $cl(E)$ of a subset $E \subset V$ is the set of vertices such that $cl(E) = E \cap bd(E)$. Finally, a probability measure follows the global Markov property, relative to G , if for any triple (P, Q, S) of disjoint subsets of V such that S separates P from Q in G so that $P \perp\!\!\!\perp Q | S$. Because the global Markov property implies the local, which in turn implies the pairwise Markov property, it is the strongest of the three. A probability distribution

P on a discrete sample space with strictly positive density satisfies the pairwise Markov property if and only if it factorizes (Lauritzen, 1996). Thus an undirected graph automatically satisfies the pairwise Markov property.

The joint densities associated to a directed acyclic graph and a chain graph factorize according to their moral graph respectively. In this case, the probability distributions follow the strongest global Markov property, which in turn implies local and pairwise Markov properties. Thus, for a directed acyclic graph and a chain graph, it is important to obtain their corresponding moral graphs as factorization of the joint densities according to these moral graphs to directly imply Markov properties.

There has been extensive research activity in developing and extending the links between graphical structures and conditional independence properties (Andersson, et al., 1995, 1997, 2001, 1993, 2006). A logical research question to explore is whether a probability distribution exists displaying the underlying properties and only the conditional properties displayed by a given graphical representation (Geiger & Pearl, 1990, 1993; Studeny & Bouckaert, 1998). It is important to note that Markov properties and conditional independence lay out one of several possible structures of an underlying graph because there may be more than one graph representing the same conditional independence relations. Over the last few decades, there has been a focus on characterizing the Markov equivalence class of graphs and nominating a natural representative of an equivalence class. After a Markov equivalence class is established and Markov properties are fulfilled, factorization of a joint probability of the attributes under study takes place uniquely, facilitating simplified computation for statistical inference.

Computational Issues in Graphical Models: Graph Structure and Statistical Learning

Graphical models have been studied extensively in order to investigate associations among discrete, continuous and mixed variables. Lauritzen and Wermuth (1989) examined properties of conditional Gaussian (CG) distributions and their applications to conditional

Gaussian regression, with emphasis on Markovian properties to attain tractable form in the subsequent analysis of these models. Due to their flexible structures and abilities to represent both structural and associative dependences, chain graph models have been studied extensively in the literature of graphical modeling (Andersson, 1997, 2006; Frydenberg, 1990). The characterization of Markov properties for chain graphs, undirected and directed acyclic graphs, has important implications to the context of factorization of underlying probability models. The Markov properties of a graph directly impact computational issues based on joint likelihood function, however, Markov properties and conditional independence may only provide one of several possible underlying graph structures. Thus, it is important to characterize Markov equivalence class of graphs and nominate a natural representative of an equivalence class.

The characterization of Markov equivalence classes has significant implications to the context of the structure of graphical models. Two graphs are Markov equivalent if they have the same Markov properties. Using results from Verma and Pearl (1991) for directed acyclic graphs, Frydenberg (1990) showed that two chain graphs are Markov equivalent if and only if they have the same skeletons, or the undirected versions, and the same complexes. A complex is a subgraph induced by a set of nodes $\{v_1, v_2, \dots, v_k\}$ with $k \geq 3$, whose edge set consists of $v_1 \rightarrow v_2$, $v_{k-1} \leftarrow v_k$, and $v_i \sim v_{i+1}$ for $2 \leq i \leq k-2$.

For a class of Markov equivalent chain graphs, a unique largest chain graph having the maximum number of undirected edges exists. The arrows of this largest graph are present in every other member of the class and thus may be considered as the representative graph of the class. There is no natural representative of an equivalence class within the class of directed acyclic graphs although it can be characterized by what is referred to as its essential graph (Andersson, et al., 1997). The natural representative of an equivalent class of chain graphs is the one with same skeleton in which an edge has an arrow, if and only if at least one member of the equivalence class has that arrow

and none has the reverse arrow (Andersson, et al., 1997).

Alternative graphical representations of conditional independence and Markov properties have been considered in the graph theory literature. Markov equivalence classes for chain graphs, undirected and directed acyclic graphs examined by Lauritzen and Wermuth (1989) and Frydenberg (1990) are referred to as LWF by Andersson, Madigan and Perlman (2001) who considered an alternative Markov property with a new semantics AMP to facilitate a direct mode of data generation (Cox, 1993; Cox & Wermuth, 1993). Andersson and Perlman (1993) showed that for AMP chain graphs, each Markov equivalence class can be uniquely represented by a single distinguished chain graph, the AMP essential graph, which plays a fundamental role in inference and model search. However, the AMP approach does not correspond to factorization of joint density in a straightforward manner; a crucial aspect for computational efficiency (Cowell, et al., 1999). Koster (1996) considered a generalization of chain graphs to reciprocal graphs and Drton (2009) showed that the block recursive Markov property of discrete chain graph models is equivalent to the global Markov property. The practical use of these models lies in developing algorithms for efficient computation characterizing running time and space complexities.

Exact and approximate inference algorithms for graphical Markov models based on independence graphs are proposed to address computational issues. Computational advancement for graphical models, particularly for the probabilistic expert systems evolved through construction of fundamental graph algorithms namely, moralization, triangulation and junction tree. The joint distribution of a graphical model can be represented and manipulated efficiently using a junction tree derived from the original graph. The junction tree algorithm starts with a moralized graph. A directed graphical model can be converted to an equivalent undirected model by the moralization process. The algorithm first selects an elimination order for all nodes and applies a triangulation operator to the moralized graph yielding a triangulated graph, then the triangulated graph creates a data structure

known as a junction tree on which a generalized message-passing algorithm can be defined (Xing, 2004). Figures 2 and 3 show an example of this process (Xing, 2004).

A junction tree possesses a key property of a running intersection, which implies that, when a node appears in any two cliques in the tree it appears in all cliques lying on the path between the two cliques. The running intersection property of the junction tree enables the joint probability distribution to be factorized as,

$$p(x) = \frac{\prod_{C_i \in \mathbf{C}_T} \psi_i(x_{C_i})}{\prod_{S_j \in \mathbf{S}_T} \phi_j(x_{S_j})}, \quad (7)$$

where $\mathbf{C}_T$ is the set of all cliques in the triangulated graph and $\mathbf{S}_T$ is the set of separators spanned by the junction tree (Xing, 2004). A message passing scheme on the junction tree updates the clique potentials $\psi(\cdot)$ and the separator potentials $\phi(\cdot)$ according to the rule,

$$\begin{aligned} \phi_j^*(x_{S_j}) &= \sum_{x_{C_i \setminus S_j}} \psi_i(x_{C_i}), \quad \psi_k^*(x_{C_k}) \\ &= \frac{\phi_j^*(x_{S_j})}{\phi_j(x_{S_j})} \psi_k(x_{C_k}), \end{aligned} \quad (8)$$

where x_{S_j} denotes the set of variables separating cliques x_{C_i} and x_{C_k} , and the message being passed from clique i to k via separator j . Running time and space complexity of the junction tree algorithm is determined by the size of the maximal clique in the triangulated graph, which is affected by the choice of elimination order that induces the triangulated graph. Tree width of a graph is the minimum of the maximal clique size among all possible triangulations. Selecting an elimination order that minimizes the maximal clique size is an NP-hard problem for arbitrary graphs. The implementation of this exact inference algorithm based on the junction tree is not efficient – or possible – for graphical models under high dimensional data. Although exact inference algorithms are simple to interpret, their implementation in high

dimensional problems becomes prohibitive due to running time and space complexities.

Computational Issues in Graphical Models: Efficient Learning/Inference Engines, Algorithms and Complexities

Approximate efficient inference algorithms, such as variational approach under a complex scenario, are considered. The approach involves converting the original optimization problem into an approximated optimization problem that is solved for an approximate

solution to the exact inference problem. Given a probability distribution $p(x|\theta)$ that factors according to a graph, the variational methods yields approximations to marginal probabilities by solving an optimization problem exploiting the underlying graphical structure (Xing, 2004). Many graphical models can be naturally viewed as an exponential family of distributions, a broad class of distributions for both discrete and continuous random variables, through the principle of maximum entropy (Wainwright & Jordan, 2008). This principle depends on a

Figure2: Moralization of a Directed Graph (Xing, 2004)

Figure3: Triangulated Graph of Figure 2 Directed Graph and Junction Tree

(a) Triangulated Graph of Figure 2 Directed Graph

(b) The Junction Tree

functional of the probability density p , absolutely continuous with respect to some measure ν . $H(p)$ is known as Shannon entropy and is defined as,

$$H(p) := - \int_x (\log p(x)) p(x) (dx) \quad (9)$$

Consider variational inference approaches for the exponential family representations of the graphical models. Approximate inference methods, such as sum-product algorithms, generalized belief propagating methods and generalized mean field inference algorithms are some of the most recent computational methodologies for graphical models in a high dimensional scenario. For variational inference, the exponential family of joint distributions determined by a collection of potential functions or sufficient statistics $\phi = \{\phi_\alpha \in C\}$ is expressed as,

$$p(\mathbf{x}|\theta) = \exp\left\{\sum_{\alpha \in C} \theta_\alpha \phi_\alpha(\mathbf{x}_{C_\alpha}) - A(\theta)\right\}, \quad (10)$$

where C is the set of cliques, C_α is the clique corresponding to the node α , $A(\theta)$ is the log partition function or cumulant function defined as

$$A(\theta) = \log \int_{X^m} \exp\left\{\sum_{\alpha \in C} \theta_\alpha \phi_\alpha(\mathbf{x}_{C_\alpha})\right\} \nu(d\mathbf{x}),$$

where X^m is a product space for m random variables. The conjugate dual function to $A(\theta)$, central to the variational principle, is defined as $A^*(\mu) := \sup_{\theta \in \Theta} \{\langle \mu, \theta \rangle - A(\theta)\}$. Here θ and μ represent canonical and mean parameters respectively of the exponential family of distributions. The conjugate dual function A^* takes the form $A^*(\mu) = -H(p_{\theta(\mu)})$, where the functional $H(\cdot)$ is defined as the Shannon entropy of the density $p_{\theta(\mu)}$ given that μ is in the interior of the set of realizable mean parameters M which is defined as,

$$M := \{\mu \in R^d \mid \exists p \text{ s.t. } E_p[\phi(X)] = \mu\}$$

(Wainwright & Jordan, 2008). Here R^d indicates number of elements to be specified in the vector

of sufficient statistics and the variational representation of the log partition function in terms its dual A^* is $A(\theta) = \sup_{\mu \in M} \{\langle \theta, \mu \rangle - A^*(\mu)\}$. Thus, under the variational representation it is necessary to maximize or minimize over the set of M as opposed to the entire parameter space Θ . The optimization problem for the variational representation of specialized graphs such as trees is computationally feasible, however, for a general structure graphical model with a large number of nodes, exact optimization becomes infeasible due to the complexity in characterizing the constraint set M and dual function $A^*(\mu)$. Approximate methods seek approximations to M and $A^*(\mu)$.

Mean-Field Methods as an Approximation to the Exact Variational Principle

In order to implement a variational inferential approach, the nature of the constraint set M and an explicit form for the dual function A^* must be known (Wainwright & Jordan, 2008); this, however, may not be easy to obtain for most practical problems. Mean field approaches permit limiting of the optimization to a subset of distributions, referred to as tractable distributions, for which both M and A^* are relatively simple to characterize (Wainwright & Jordan, 2008). For a graphical model based on a graph $G(V, E)$, the tractability can be obtained in terms of a tractable sub-graph. A sub-graph F is tractable if it is possible to carry out exact calculations on F .

A straightforward example of a tractable sub-graph is the fully disconnected sub-graph $F(V, \emptyset)$ containing all the vertices of $G(V, E)$ but none of the edges. This tractable sub-graph F leads to a product distribution for which computations are easy to carry out. However, completely disconnected sub-graphs do not capture dependencies among vertices, if any. Thus, as opposed to a fully disconnected sub-graph, an arbitrarily structured sub-graph from the given graph $G(V, E)$ is considered in generalized mean field methods. The question then becomes how to select a tractable sub-graph leading to an efficient factorization of the joint probability distribution so that feasible solution set M and the optimizing function A^* can be

characterized with less intensive computational and mathematical background. In addition, a generalized version of the mean field methods to the context of chain graph models is of great importance for practical problems in numerous scientific fields.

Discussion and Future Direction Statistical Learning on the Underlying Graph Structure from Empirical Data

In order to implement chain graph models to study relationships among stochastic variables in empirical data, consider the exponential family of probability distributions as the distribution of the random variables associated to the nodes of a graphical model. It is relatively straightforward to write the joint probability distribution of a set of discrete random variables utilizing the factorization under a given graphical structure. The factorization of joint distribution of continuous random variables representing the nodes of an underlying graph requires attention and a general framework for the factorization scheme of joint probability distributions of both discrete and continuous random variables using established graph theory properties is of interest for simplified computation.

Gaussian graphical models for continuous variables and the graphical counterpart of log-linear models for discrete attributes are proposed and implemented for empirical model building. For a large volume of attributes, as in biological network data, such as gene-gene interaction networks, gene-protein interaction networks and transcription regulatory networks, as well as network data in other scientific and social science fields, these methods can be computationally prohibitive.

The variational inference approach based on the mean parameterization of the exponential family of distributions and their mathematical properties, such as, convexity and conjugate duality is an efficient inference approach to graphical models. Implementation of these algorithms and complexity are of interest in the contexts of high dimensional gene-gene interaction networks, gene-protein interaction networks and transcription regulatory network data. Structural properties such as connectivity and existence of specific

substructures in the graphical models are of specific interest. It is necessary to identify Markov equivalence classes in order to narrow down possible representations of same conditional independence by many graphical structures.

In particular, this investigation considers when a chain graph G_{ch} is Markov equivalent to some unique undirected graph G_U , decomposable undirected graph and to some directed acyclic graph G_D . The directed acyclic graph models provide a convenient recursive factorization of the joint probability. The likelihood function factorizes and it is possible to implement maximum likelihood methods for estimation of model parameters. These tractable features are also available for decomposable undirected graphs, which are Markov equivalent to some directed acyclic graphs (Andersson, et al., 1997). For a directed acyclic Markov model G_D , the joint density factorizes as

$$p(x | \theta) = \prod_{v \in V} p(x_v | x_{pa(v)}, \theta^{v, x_{pa(v)}}) \quad (11)$$

where $\theta^{v, x_{pa(v)}}$ is the minimal function of the overall parameter θ for the distribution determining the conditional distribution of $X_v | X_{pa(v)} = x_{pa(v)}$. For a complete case, each factor in the likelihood is maximized separately to attain a maximum likelihood estimate of $\theta^{v, x_{pa(v)}}$. For an incomplete case consider the expectation maximization (EM) algorithm. Let $f(x | \theta)$ denote the density function of a random variable X that is incomplete except one known function, $Y = g(X)$. Given an initial estimate θ , the E-step requires the current expected value of the log-likelihood function $Q(\theta | \theta) = E_{\theta} \{ \log f(X | \theta) | g(X) = y \}$. The M-step maximizes Q over θ' yielding the next estimate. The algorithm alternates between these two steps until convergence is attained. The evidence propagation or message-passing on the junction tree can be exploited to perform the E-step of an EM algorithm for a discrete directed acyclic graph model with missing observations (Lauritzen, 1995). Gradient-descent search near the maximum can be considered to speed up the convergence of an expectation-maximum (EM)

algorithm in a graphical model with incomplete data.

A general factorization scheme for the joint probability distribution in the exponential family enabling tractability in subsequent statistical computation sets the foundation for efficient computation in graphical models. Searching within a collection of candidate models for one or more graphical structures consistent with the conditional independence relationships suggested by data follows; the point is to assess the adequacy of a single candidate graphical model as the so-called true objective process of generation of empirical data. In order to narrow down the possible high dimension of the space of the graphical structures, the Markov equivalence classes of graphical structures, identification of a unique graphical model as the probability model and checking identifiability of the model parameters are essential. Either the likelihood-based or the Bayesian methods can be implemented to address the estimation and model search problem. Complete case data are addressed through maximum likelihood estimation or a Bayesian updating scheme; incomplete case data are addressed through the EM algorithm coupled with gradient search methods for estimation using likelihood- and sampling- based methods using a Bayesian approach respectively.

Model Selection, Diagnostics and Checking Models against Data

A Markov equivalence class insures only proper graphical structure. The properties of the joint probability distribution of the variables must be inferred from the graphical structure and the conditional independence relationships suggested by the empirical data. According to the semantics of machine learning and data mining, unsupervised learning methods for model selection, diagnostics and model checking against data can be employed. In low dimensional problems with a number of variables $p \ll q$, effective nonparametric methods are used for density estimation solely from the data (Silverman, 1986). However, these methods are not applicable in high-dimension problems due to the curse of dimensionality.

Testable hypotheses, based on prior knowledge and expert opinion in the scientific

field along with corresponding testing principles should be developed to address the graphical model selection problem. Efficient computational algorithms along with running time and space complexities must be formulated. Diagnostics in statistical modeling address outlier detection problems and development of robust methods against outliers. Outlier identification in high dimensional problems is an active research area where robust principal component analysis, k - nearest neighbor, local outlier factor and other distance and density based methods are commonly used. Future research interests should center on addressing these important statistical problems for high-dimensional data.

Conclusion

Graphical models originated as the marriage between graph and probability theories and are appealing methods for studying conditional (in)dependences among a large number of attributes in many scientific fields. Markov properties of various graphical models, directed, undirected and more general chain graph models, lead to efficient factorization of joint probability distributions of multivariate random variables. An explicit form of a joint distribution may not be known for many random variables, except some arbitrary dependence structure.

Graphical modeling is an efficient tool for studying dependence structure among an arbitrary number of random variables without specifying their joint distribution. This article described essential properties of graphical models that lead to factorization of a joint distribution. An exponential family representation of graphical models was demonstrated for a broad class of distributions of discrete and continuous random variables. Exponential family representation is essential for formulating approximate inference algorithms such as mean field algorithms. It was also indicated that studies regarding unique graph structure through a Markov equivalence class of graphs for specialized undirected, directed and general chain graphs is an area for future research. Finally, a graphical model derived from a unique graph structure illuminated the relationship among the attributes under study.

References

- Andersson, S. A., Madigan, D., & Perlman, M. D. (1995). *A characterization of Markov equivalence classes for acyclic digraphs: Technical report 287*. University of Washington, Seattle: Department of Statistics.
- Andersson, S. A., Madigan, D., & Perlman, M. D. (1997). On the Markov equivalence of chain graphs, undirected graphs and acyclic digraphs. *Scandinavian Journal of Statistics*, 24, 81-102.
- Andersson, S. A., Madigan, D., & Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25, 505-541.
- Andersson, S. A., Madigan, D., & Perlman, M. D. (2001). Alternative Markov Properties for chain graphs. *The Scandinavian Journal of Statistics*, 28, 33-85.
- Andersson, S. A., & Perlman, M. D. (1993). Lattice models for conditional independence in a multivariate normal distribution. *The Annals of Statistics*, 21, 1318-58.
- Andersson, S. A., & Perlman, M. D. (2006). Characterizing Markov equivalence classes for AMP chain graph models. *The Annals of Statistics*, 34, 939-972.
- Breiman, L. (1993). Better subset selection using the non-negative garotte. *Technical Report*. University of California, Berkeley.
- Chipman, H. A., George, E. I., & McCulloch, R. (2009). *Bayesian additive regression trees*. Preprint, NI09002-SCH. Cambridge, MA: Isaac Newton Institute.
- Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statistical Science*, 22, 1-43.
- Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. New York: Springer.
- Cox, D. R. (1993). Causality and graphical models. *Bulletin of the International Statistical Institute, Proceedings of the 49th Session*, 1, 363-372.
- Cox, D. R., & Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statistical Science*, 8, 204-218; 247-277.
- Drton, M. (2009). Discrete chain graph models. *Bernoulli*, 15(3), 736-753.
- Edwards, D. (2000) *Introduction to graphical modeling*. New York, NY: Springer-Verlag, Inc.
- Frydenberg, M. (1990). The Chain Graph Markov property. *Scandinavian Journal of Statistics*, 17, 333-353.
- Frydenberg, M. (1990). Marginalization and collapsibility in graphical interaction models. *The Annals of Statistics*, 18, 790-805.
- Geiger, D., & Pearl, J. (1990). On the logic of causal models. In *Uncertainty in Artificial Intelligence*, R. D. Shachter, T. S. Levit, L. N. Kanal, & J.F. Lemmer, Eds., 4, 3-14. Amsterdam, The Netherlands: North-Holland.
- Geiger, D., & Pearl, J. (1993). Logical and algorithmic properties of conditional independence and graphical models. *The Annals of Statistics*, 21, 2001-2021.
- Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. London, UK: Chapman & Hall/CRC.
- Koster, J. T. A. (1996). Markov properties of nonrecursive causal models. *The Annals of Statistics*, 24, 2148-2177.
- Lauritzen, S. L. (1996). *Graphical models*. Oxford, England: Oxford Science Publications.
- Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19, 191-201.
- Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17, 31-57.
- Neapolitan, R. E. (1990). *Probabilistic reasoning in expert systems*. New York, NY: John Wiley and Sons.
- Neapolitan, R. E. (2004). *Learning Bayesian networks*. Upper Saddle River, NJ: Prentice Hall.
- Neapolitan, R. E. (2009). *Probabilistic methods for bioinformatics*. Burlington, MA: Morgan Kaufmann.

Ravikumar, P., Liu, H., Lafferty, J., & Wasserman, L. (2007). SpAM: sparse additive models. In *Advances in neural information processing systems*, J. C. Platt, D. Koller, Y. Singer & S. Roweis, Eds., 1201-1208. Cambridge, MA: MIT Press.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. London, England: Chapman and Hall.

Studeny, M., & Bouckaert, R. R. (1998). On chain graph models for description of conditional independence structures. *The Annals of Statistics*, 26, 1434-1495.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

Verma, T., & Pearl, J. (1991). Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence*, R. D. Shachter, T. S. Levit, L. N. Kanal, & J.F. Lemmer, Eds., 255-268. Amsterdam, The Netherlands: North-Holland.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1, 1-305.

Wermuth, N., & Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, 70, 537-552.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York, NY: John Wiley and Sons.

Xing, P. (2004). *Probabilistic graphical models and algorithms for genomic analysis*. PhD. Dissertation, Department of Computer Science, University of California, Berkeley.

Small-to-Medium Enterprises and Economic Growth: A Comparative Study of Clustering Techniques

Karim K. Mardaneh
University of Ballarat,
Mount Helen, Australia

Small-to-medium enterprises (SMEs) in regional (non-metropolitan) areas are considered when economic planning may require large data sets and sophisticated clustering techniques. The economic growth of regional areas was investigated using four clustering algorithms. Empirical analysis demonstrated that the modified global k -means algorithm outperformed other algorithms.

Key words: Clustering, k -means, Ward's clustering, firm size, industry structure, regional economy.

Introduction

Clustering techniques are compared by examining the relationship between industry structure and business size with economic growth using Australian regional areas (non-metropolitan) data. Pagano (2003) examined firm size and industry structure; however, the study did not consider in combination the role of both industry structure and size of business in economic growth. This study uses four clustering techniques on statistical local area (SLA) regions to examine the performance of these clustering methods on small-to-medium enterprises (SMEs) data sets. Researchers such as Beer and Clower (2009) have used clustering techniques for pattern recognition; however, there is a gap in the literature in terms of applying clustering methods to SMEs related problems.

Data mining facilitates the identification of useful information within data reservoirs and involves the application of discovery algorithms

to the data. Cluster analysis is an important data mining task (Mardaneh, 2007). Cluster analysis has been used by contemporary researchers when the number of observations in a particular field is fairly large (Freestone, Murphy and Jenner, 2003). This study adopts cluster analysis and uses four methods of clustering algorithms: Ward's (Ward, 1963), the k -means (Hartigan & Wong, 1979), global k -means (Likas, Vlassis & Verbeek, 2003), and the modified global k -means (Bagirov, 2008; Bagirov & Mardaneh, 2006). These algorithms are employed to cluster SLAs based on industry structure and size of the businesses within those areas and to compare the function of the algorithms to identify a clustering algorithm that is most suitable for clustering SMEs data. This study addresses the gap in understanding the combined role of industry structure and size of business in economic growth, as well as the cluster analysis of the SMEs data sets.

Literature Review

Understanding economic growth requires a thorough consideration of the role of industry structure and the size of business (micro, small, medium or large). Regions with an industry structure that enables wealth-creating initiatives will have a better economic condition (Delgado, et al., 2010). In addition, the distribution of a region's economic activity across industries is considered to be a major determinant of the resilience of its economy (Australian Government Department of

Karim Ken Marandeh is a Lecturer and Post Doctoral Fellow in the Business School. His research interests include regional economics, economy of small to medium size enterprises, data mining and its application in economics and business problems. Email him at: k.mardaneh@ballarat.edu.au.

Transport and Regional Services, 2003). Previous studies in this area mainly focus on formation and growth (Dobbs & Hamilton, 2007; Mueller, et al., 2008; Hudson, et al., 2001; Beugelsdijk, 2007; Sierdjan, 2007; Koster, 2007; Armington & ACS, 2002; Pagano & Schivardi, 2000, Dejardin & Fritsch, 2010), as well as organizational attitude to change, and success and failure (Walker & Brown, 2004; Agarwal & Audretsch, 2001; Gray, 2002; Feser, et al., 2008; Dejardin & Fritsch, 2010). A few studies have investigated industry structure and size of business (Okamuro, 2006, Okamuro & Hobayashi, 2006; Pagano, 2003; Pagano & Schivardi, 2000); however, these studies did not identify the drivers of economic growth in relation to those factors.

Clustering, or cluster analysis, is a challenging problem for which different algorithms have been proposed. Cluster analysis addresses the problem of organizing a collection of patterns or objects into clusters based on similarity so that objects in the same cluster are in some way more similar to each other than to those in other clusters (Bagirov, 2008; Bagirov & Mardaneh, 2006). Beer and Maude (1995) used cluster analysis to examine changes in economic functions of towns, and Smith (1965) used clustering in the study of economic functions of Australian regional towns. In this study the clustering technique is used to collect SLAs into clusters so that SLA regions within a cluster are similar to each other and are different from regions in the other clusters.

Clustering methods in general have been used in business and economics disciplines. Ward's clustering method has been widely used in consumer behavior (Greeno, Sommers & Kernan, 1973; Kernan & Bruce, 1972), marketing and economics studies (Eliashberg, Lilien & Kim, 1995; Blin & Cohen, 1997; Doyle & Saunders, 1985). Ward's clustering in particular has been used to study Australian regional economic development (Beer & Maude, 1995; Beer & Clower, 2009; Sorensen & Weinand, 1991), urbanization in Australian economy (Freestone, et al., 2003), and marketing themes and strategies (Ho & Hung, 2008; Wong & Saunders, 1993). Unlike the Ward's method, the *k*-means algorithms have not been widely used within these disciplines.

The *k*-means algorithms have been mainly used in information technology and data mining studies and in a few marketing studies (Calantone & Sawyer, 1978; Moriarty & Venkatesan, 1978; Schaninger, Lessig & Panton, 1980). The *k*-means algorithm has only recently been used in regional economics studies (Mardaneh, 2012).

This study explores whether the *k*-means algorithm and its variations could provide a better tool for regional economics studies than the Ward's clustering algorithm. The framework of the study is based on SMEs, the two variables (industry structure and business size) and the comparative experiment of the four algorithms. A more efficient algorithm that better clusters the SMEs data could help advance the understanding of industry structure and size of business (SMEs) which, in turn, could provide valuable information regarding regional economic growth.

Methodology

Using regional Australian data this study examines the influence of the industry structure and size of businesses on the economic growth of SLAs. To measure growth, individual weekly income is used as a proxy for economic growth and assumes that SLAs with more people in \$1,000-\$1,999 and \$2,000 and over income per week must enjoy a particular industry structure and business sizes. To investigate this, SLAs based on industry structure and three business sizes (micro, small, medium) are clustered. Clustering is conducted three times, once for each size of business, using the *k*-means, global *k*-means, modified global *k*-means and Ward's clustering algorithms. Results are compared to identify the clustering algorithm that is most suitable for clustering the SMEs data.

Data for this study is obtained from the Australian Bureau of Statistics (ABS, 2007) and uses information from the *Counts of Australian Businesses, including Entries and Exits, June 2003-June 2007*, which includes *Businesses by Industry Division by SLA by Employment size ranges*. This is provided as categories of data for businesses by industry division (see Appendix A for the list of industries). The data exhibits sixteen industry types and the number of employees at each SLA based on business size.

The ABS classifies size of businesses as micro business (1-4 employees), small business (5-19 employees), medium business (20-199 employees) and large business (200 and over employees). This classification is maintained herein, however, this study does not include large businesses (200 and over employees) because the relevant data were too sparse. For the same reason the 'electricity, gas and water supply' industry is excluded from the analysis. Because this study focuses on regional geographical areas - and due to the fact that the industry structure and number of business sizes in regional areas are very different from metropolitan areas - metropolitan data is excluded; this avoids skewness in analysis. After removing metropolitan SLAs and outliers (extreme values in data set) 661 regional SLAs were included in the analysis.

The percentage of people in two weekly income levels, \$1,000-\$1,999 and \$2,000 and over, are considered per SLA. The median of the percentage for each income level is calculated across all SLAs (11.8% and 1.9% for each income level, respectively). SLAs above median within both income levels are considered as SLAs having a higher level of economic growth and are labeled as category 1; the remaining SLAs are considered as SLAs with a lower level of economic growth and are labeled as category 2.

Samples in the data are comprised of SLAs under three business sizes (1-4, 5-19, 20-199) and fifteen industry types which form the data set. To identify the industry type(s) and business size(s) with higher or lower levels of contribution to the economic growth of a SLA (which allocates a SLA to categories 1 or 2) clustering analysis was conducted using three SMEs data sets (see Tables 1-3). For this, the k -means, global k -means, modified global k -means and Ward's clustering algorithms were applied (see Tables 4-6).

Clustering Algorithms

Clustering algorithms can be used to analyze large data sets comprising a myriad of economic and social variables. They seek to group samples with similar characteristics and ensure maximum statistical separation from other contrasting clusters. This is a process of

pattern recognition which simplifies understanding of large data sets. In one classification, clustering algorithms are classified as hierarchical or iterative algorithms. Hierarchical methods begin with a set of clusters and place each sample in an individual cluster. Clusters are then successively merged to form a hierarchy of clusters (Guha, et al., 2001). Iterative methods start by dividing observations into some predetermined number of clusters. Observations are then reassigned to clusters until some decision rule terminates the process (Punj & Stewart, 1983). Ward's clustering algorithm is hierarchical, while the k -means and its variations are iterative.

Ward's Algorithm

Ward's algorithm seeks to group a set of n members, which are called subsets or groups in relation to an objective function value. The method seeks to unite two of the n subsets to reduce the number of subsets to $n-1$ in a way that minimizes the change in the objective function's value. The $n-1$ resulting subsets are examined to determine if a third member should be grouped with the first pair. If necessary this procedure can be continued until all n members of the original array are in one group (Ward, 1963).

The k -means Algorithm

The k -means algorithm considers each sample (SLAs in this study) as a point in n -dimensional space (R^n) and chooses k centers (also called centroids) and assigns each point to the cluster nearest the center. The center is the average of all points in the cluster, that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. The k -means algorithm is an efficient clustering algorithm, but it is sensitive to the choice of starting points (Bagirov, 2008).

The Global K -means Algorithm

The global k -means algorithm was proposed to improve global search properties of k -means algorithms. The global k -means algorithm (Likas, et al., 2003) computes clusters successively. At the first iteration of this algorithm the centroid of a set A is computed

and, in order to compute k -partition of the k^{th} iteration, the algorithm uses centers of $k-1$ clusters from the previous iteration (Likas, et al., 2003).

The Modified Global K -means Algorithm

The modified global k means algorithm computes clusters incrementally and, to compute the k -partition of a data set, it uses $k-1$ cluster centers from the previous iteration. An important step in this algorithm is the computation of a starting point for the k^{th} cluster center. This starting point is computed by minimizing the so-called auxiliary cluster function. (Bagirov, 2008; Bagirov & Mardaneh, 2006)

Empirical studies of the performance of clustering algorithms (Punj & Stewart, 1983) suggest that one of the iterative clustering methods (e.g., k -means clustering) is preferable to hierarchical methods (e.g., Ward's clustering). The k -means appears to be more efficient (Mezzich, 1978; Milligan, 1980; Bayne, et al., 1980) if a non-random starting point is specified. When a clustering algorithm includes more and more observations its performance tends to deteriorate. This effect may be the result of outliers entering into the solution. The k -means appears to be more robust than hierarchical methods with respect to the presence of outliers. Results from this study suggest that the more efficient version of the k -means algorithm (modified global k -means) may better cluster SMEs data and could help with further understanding of industry structure and the size of business in regional economics studies.

Results

The analysis clustered SLAs based on industry type and three business sizes (1-4, 5-19, 20-199). Industry, cluster category and the cluster centroids values are reported in Tables 1-3; industries are reported in the tables only if the difference between cluster centroids values in cluster category 1 and 2 for a particular industry is 0.1 or more. In addition, industry type and size of business (variables) with higher cluster centroids value in cluster category 1 are considered as variables with a higher level of contribution to economic growth. Industry type and size of business with higher cluster centroids

value in cluster category 2 are considered as variables with a lower level of contribution to economic growth.

As shown in Tables 1-3, the construction, retail trade and personal and other services industries indicate a higher level of contribution to economic growth in all three firm sizes. By contrast, the agriculture, forestry and fishing and wholesale and communication services industries show a lower level of contribution to economic growth in all three firm sizes.

The property and business services industry shows a higher level of contribution to economic growth for both firm sizes 1-4 and 5-19; however, this industry shows a lower level of contribution to economic growth in firms sized 20-199. The cultural and recreational services industry shows a higher level of contribution in both 1-4 and 20-199 sized firms, but shows a lower level of contribution for those sized 5-19. The transport and storage industry shows a higher level of contribution for both 5-19 and 20-199 sized firms; however, it shows a lower level of contribution for 1-4 sized firms.

The health and community services industry shows a higher level of contribution for firms sized 5-19, but shows a lower level of contribution for firms sized 1-4. The finance and insurance industry shows a higher level of contribution only for larger firms, size 20-199. The accommodation, cafes and restaurants industry shows a higher level of contribution for sized 20-199 firms; however, it shows a lower level of contribution for both of the other two sizes. The mining and manufacturing industries both show a lower level of contribution for 1-4 and 20-199 sized firms.

By applying clustering analysis, this study sought to identify the most efficient algorithm for clustering SMEs data. For this, the objective function value and the CPU time spent by each algorithm for clustering were calculated. Clustering was conducted for 2, 5, 10, 15, and 20 cluster numbers for comparison. The analyses in this study were conducted using an Intel Core 2 Duo, 2.99 GHz, PC. Tables 4- 6 show the number of clusters (N), values of the objective function ($f \times 10^5$) and CPU time spent for the

Table 1: Higher/Lower Level of Industry Contribution in Economic Growth;
Firm Size 1-4

Industry	Cluster Category	
	1	2
	Cluster Centroids	
Higher Level of Contribution to Economic Growth		
• Construction	17.42	13.62
• Retail Trade	14.33	11.98
• Property and Business Services	12.48	11.55
• Personal and other Services	3.68	3.55
• Cultural and Recreational Services	1.61	1.41
Lower Level of Contribution to Economic Growth		
• Mining	0.51	0.76
• Communication Services	1.54	1.90
• Wholesale	3.81	4.09
• Accommodation, Cafes and Restaurants	3.53	4.50
• Health and Community Services	4.79	4.92
• Manufacturing	4.98	5.13
• Transport and Storage	5.84	6.10
• Agriculture, Forestry and Fishing	21.32	26.81

Table 2: Higher/Lower Level of Industry Contribution in Economic Growth;
Firm Size 5-19

Industry	Cluster Category	
	1	2
	Cluster Centroids	
Higher Level of Contribution to Economic Growth		
• Retail Trade	17.64	14.88
• Construction	12.41	8.35
• Property and Business Services	11.43	10.83
• Transport and Storage	5.07	4.87
• Health and Community Services	5.01	4.88
• Personal and other Services	3.57	3.40
Lower Level of Contribution to Economic Growth		
• Communication Services	0.60	0.77
• Cultural and Recreational Services	1.70	1.88
• Wholesale	3.99	5.10
• Accommodation, Cafes and Restaurants	7.50	7.69
• Agriculture, Forestry and Fishing	20.99	27.97

Table 3: Higher/Lower Level of Industry Contribution in Economic Growth;
Firm Size 20-199

Industry	Cluster Category	
	1	2
	Cluster Centroids	
Higher Level of Contribution to Economic Growth		
• Retail Trade	15.89	13.69
• Accommodation, Cafes and Restaurants	11.73	11.03
• Construction	10.17	5.5
• Transport and Storage	6.14	4.73
• Cultural and Recreational Services	3.50	3.26
• Finance and Insurance	2.00	1.39
• Personal and other Services	1.50	1.30
Lower Level of Contribution to Economic Growth		
• Communication Services	0.41	0.98
• Mining	0.66	1.19
• Wholesale	4.39	5.72
• Manufacturing	8.80	10.11
• Property and Business Services	10.11	11.35
• Agriculture, Forestry and Fishing	16.91	20.93

analysis (t) for the multi-start k -means (MSKM), global k -means (GKM), modified global k -means (MGKM), and Ward's (WARD) clustering algorithms. Results from the analysis, including the objective function values and CPU time spent for the calculation by each algorithm, are shown in Tables 4-6.

Algorithm Performance

Results presented in Table 4 show that MGKM algorithm outperforms both the MSKM and GKM when the number of clusters $N \geq 10$. Regardless of the number of clusters, the MGKM outperforms WARD and WARD gives the worst results compared to all other algorithms. The GKM requires less CPU time; however, its solutions are not better. MGKM requires more CPU time, particularly when the number of clusters increases ($N \geq 10$). Similarly CPU time for MSKM and GKM increases as the

number of clusters N increases. CPU time for WARD is nearly constant for any cluster number N because it is a hierarchical algorithm and, unlike the other three algorithms, it does not go through iterations.

Results in Table 5 show that the MGKM algorithm outperforms both MSKM and GKM when the number of clusters $N \geq 10$. With any number of clusters MGKM outperforms WARD. MGKM requires more CPU time particularly when the number of clusters increases ($N > 5$). Similarly CPU time for MSKM and GKM increases as the number of clusters N increases. CPU time for WARD is almost constant for any cluster number N . Table 6 shows that in some cases, for example, $N = 2, 15$, MSKM performed slightly better than MGKM, however, the difference in performance is minimal. With any number of clusters MGKM outperforms WARD. MGKM required more CPU time for all cluster

Table 4: Data Set 1* - Comparative Values for Algorithms;
Firm Size 1-4

N	MSKM		GKM		MGKM		WARD	
	$f \times 10^5$	t	$f \times 10^5$	t	$f \times 10^5$	t	$f \times 10^5$	t
2	7.582	0.01	7.582	0.01	7.582	0.03	8.274	0.20
5	5.018	0.07	5.018	0.07	5.018	0.12	5.471	0.18
10	3.747	0.12	3.721	0.17	3.721	0.29	4.054	0.18
15	3.111	0.20	3.044	0.26	3.025	0.45	3.268	0.18
20	2.617	0.32	2.542	0.39	2.549	0.57	2.759	0.18

*Data Set 1 includes micro businesses with 1-4 employees across 15 industry types.

Table 5: Data Set 2* - Comparative Values for Algorithms;
Firm Size 5-19

N	MSKM		GKM		MGKM		WARD	
	$f \times 10^5$	t	$f \times 10^5$	t	$f \times 10^5$	t	$f \times 10^5$	t
2	8.721	0.00	8.721	0.01	8.721	0.03	9.122	0.18
5	5.955	0.04	5.944	0.07	5.944	0.12	6.376	0.18
10	4.331	0.10	4.355	0.18	4.341	0.29	4.705	0.18
15	3.609	0.23	3.605	0.28	3.570	0.45	3.888	0.20
20	3.208	0.31	3.201	0.39	3.133	0.62	3.413	0.18

*Data Set 2 includes small businesses with 5-19 employees across 15 industry types.

Table 6: Data Set 3* - Comparative Values for Algorithms;
Firm Size 20-199

N	MSKM		GKM		MGKM		WARD	
	$f \times 10^5$	t	$f \times 10^5$	t	$f \times 10^5$	t	$f \times 10^5$	t
2	15.929	0.01	15.930	0.01	15.930	0.03	16.453	0.18
5	11.488	0.03	11.058	0.09	11.058	0.12	12.164	0.17
10	7.811	0.10	7.811	0.18	7.814	0.28	8.818	0.18
15	6.324	0.15	6.336	0.28	6.345	0.43	7.029	0.18
20	5.607	0.32	5.494	0.35	5.513	0.60	6.062	0.18

*Data Set 3 includes medium businesses with 20-199 employees across 15 industry types.

numbers N . CPU time for GKM and MGKM increased as the number of clusters N increased. CPU time for WARD is nearly constant for any cluster number N .

Identifying a clustering algorithm that could help with more efficient cluster analyses of SMEs data is important. A more efficient clustering algorithm may help provide a more accurate and precise grouping of the data points (in this study, geographical areas) based on their similarity. This, in turn, will help with identifying shared characteristics between members (data points) of a cluster. Understanding these characteristics provides a diagnostic of the factors that generate those characteristics. As this study shows, such an understanding can help with identifying the role that each combined industry and business size could play in the economic growth or decline of geographical areas and also whether they have higher or lower contributions to the economic growth of an area.

Conclusion

Cluster analysis revealed clusters of industries associated with industry structure and size of business. This study presented numerical results from three data sets. The results clearly show that the modified global k -means algorithm is more efficient for solving clustering problems in SMEs data sets; this algorithm outperforms multi-start k -means, global k -means and Ward's clustering algorithms. The modified global k -means algorithm, however, requires more computational efforts than the global k -means algorithm, but is the most promising among all tested algorithms.

The findings from this study provide an improved method for clustering using a more efficient algorithm and, as a result, provide a better understanding of industry structure and size of businesses in regional areas. These findings have policy implications for future economic planning and focus on SMEs for regional areas and will provide paths in identifying significant factors that require further investigation using qualitative methods to ascertain the importance of the clusters and their relationship to SMEs.

References

- Agarwal, R., & Audretsch, D. B. (2001). Does Entry Size Matter? The impact of the life cycle and technology on firm survival. *The Journal of Industrial Economics*, 49, 21-43.
- Armington, C., & Acs, Z. J. (2002). The determinants of regional variation in new firm formation. *Regional Studies*, 36, 33-45.
- Australian Bureau of Statistics. (2007). *Counts of Australian businesses, including entries and exits, June 2003 to June 2007*, Cat. No. 8165.0. Canberra: ABS.
- Australian Government Department of Transport and Regional Services. (2003). *Information Paper 49: Focus on Regions: No.1: Industry structure*, 1-67.
- Bagirov, A. M. (2008). Modified global k -means algorithm for minimum sum-of-squares clustering problems. *Pattern Recognition*, 41, 3192-3199.
- Bagirov, A. M., & Mardaneh, K. (2006). Modified global k -means algorithm for clustering in gene expression data sets. In *Proceedings of 2006 Workshop on intelligent systems for bioinformatics (WISB 2006)*, M. Boden. & T. L. Bailey (Eds.), 23-28. Hobart: Australian Computer Society (ACS).
- Bayne, C. K., Beauchamp, J. J., Begovich C. L., & Kane, V. E. (1980). Monte Carlo comparisons of selected clustering procedures. *Pattern Recognition*, 12, 51-62.
- Beer, A., & Clower, T. (2009). Specialization and growth: Evidence from Australia's regional cities. *Urban Studies*, 46, 369-388.
- Beer, A., & Clower, T. (2009). Specialization and growth: Evidence from Australia's regional cities. *Urban Studies*, 46, 369-388.
- Beer, A., & Maude, A. (1995). Regional cities in the Australian urban system, 1961-1991. *Urban Policy and Research*, 13, 135-148.
- Beugelsdijk, S. (2007). Entrepreneurial culture, regional innovativeness and economic growth. *Journal of Evolutionary Economics*, 17, 187-210.
- Blin, J. M., & Cohen, C. (1997). Technological similarity and aggregation in input-output systems: A cluster-analytic approach. *The Review of Economics and Statistics*, 59, 82-91.

- Calantone, R. J., & Sawyer, A. G. (1978). The stability of benefit segments. *Journal of Marketing Research*, 15, 395-404.
- Dejardin, M., & Fritsch, M. (2010). Entrepreneurial dynamics and regional growth. *Small Business Economics*, 36, 377-382.
- Delgado, M., Porter, M. E., & Stern, S. (2010). Clusters and entrepreneurship. *Journal of Economic Geography*, 10, 495-518.
- Dobbs, M., & Hamilton, R. T. (2007). Small business growth: recent evidence and new directions. *International Journal of Entrepreneurial Behaviour & Research*, 13, 296-322.
- Doyle, P., & Saunders, J. (1985). Market segmentation and positioning in specialized industrial markets. *Journal of Marketing*, 24-32.
- Eliashberg, J., Lilien, G. L., & Kim, N. (1995). Searching for generalization in business marketing negotiations. *Marketing Science*, 14, 47-60.
- Feser, E., Renski, H., & Goldstein, H. (2008). Clusters and economic development outcomes: An analysis of the link between clustering and industry growth. *Economic Development Quarterly*, 22, 324-344.
- Freestone, R., Murphy, P., & Jenner, A. (2003). The functions of Australian towns revisited. *Tijdschrift voor Economische en Sociale Geografie*, 94, 188-204.
- Gray, C. (2002). Entrepreneurship, resistance to change and growth in small firms. *Journal of Small Business and Enterprise Development*, 9, 61-72.
- Greeno, D. W., Sommers, M. S., & Kernan, J. B. (1973). Personality and implicit behavior patterns. *Journal of Marketing Research*, 10, 63-69.
- Guha, S., Rastogi, R., & Shim, K. (2001). Cure: An efficient clustering algorithm for large databases. *Information Systems*, 26, 35-58.
- Hartigan, J. A., & Wong, M. A. (1979). A *k*-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28, 100-108.
- Ho, H-F., & Hung, C-C. (2008). Marketing mix formulation for higher education: An integrated analysis employing analytic hierarchy process, cluster analysis and correspondence analysis. *International Journal of Educational Management*, 22, 328-340.
- Hudson, M., Smart, A., & Bourne, M. (2001). Theory and practice in SME performance measurement systems. *International Journal of Operations & Production Management*, 21, 1096-1115.
- Kernan, J. B., & Bruce, G. D. (1972). The socioeconomic structure of an urban area. *Journal of Marketing Research*, 9, 15-18.
- Koster, S. (2007). The entrepreneurial and replication function of new firm formation. *Tijdschrift voor Economische en Sociale Geografie*, 98, 667-674.
- Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global *k*-means clustering algorithm. *Pattern Recognition*, 36, 451-461.
- Mardaneh, K. (2007). *Optimization based clustering and classification algorithms in analysis of microarray gene expression data sets*. Ph.D. thesis, School of Information Technology and Mathematical Sciences, University of Ballarat.
- Mardaneh, K. (2012). A Study of population change via clustering of Australian regional areas: An optimisation approach. *Australasian Journal of Regional Studies*, 1, 257-280.
- Mezzich, J. E. (1978). Evaluation clustering methods for psychiatric diagnosis. *Biological Psychiatry*, 13, 265-281.
- Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45, 325-342.
- Moriarty, M., & Venkatesan, M. (1978). Concept evaluation and market segmentation. *Journal of Marketing*, 42, 82-86.
- Mueller, P., Stel, A. V., & Storey, D. J. (2008). The effects of new firm formation on regional development over time: The case of Great Britain. *Small Business Economics*, 30, 59-71.

SMES AND ECONOMIC GROWTH: COMPARATIVE CLUSTERING TECHNIQUES

Okamuro, H. (2006). Determinants of R&D activities by start-up firms: Evidence from Japan. *Institute of Economic Research, Hitotsubashi University*, Working Papers Series, No. 2006-2.

Okamuro, H., & Hobayashi, N. (2006). The impact of regional factors on the start-up ratio in Japan. *Journal of Small Business Management*, 44, 310-313.

Pagano, P. (2003). Firm size distribution and growth. *Scandinavian Journal of Economics*, 105, 255-274.

Pagano, P., & Schivardi, F. (2000). *Firm size distribution and growth*, mimeo. Retrieved December 10, 2010 from: <http://digilander.libero.it/fshivardi>.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 10, 134-148.

Schaninger, C. M., Lessig, V. P., & Panton, D. B. (1980). The complementary use of multivariate procedures to investigate nonlinear and interactive relationship between personality and product usage. *Journal of Marketing Research*, 17, 119-124.

Sierdjan, K. (2007). The entrepreneurial and replication function of new firm formation. *Tijdschrift voor Economische en Sociale Geografie*, 98, 667-674.

Smith, R. H. T. (1965). Method and purpose in functional town classification. *Annals of the Association of American Geographers*, 55, 539-548.

Sorensen, T., & Weinand, H. (1991). Regional well-being in Australia revisited. *Australian Geographical Studies*, 29, 42-70.

Walker, E., & Brown, A. (2004). What success factors are important to small business owners? *International Small Business Journal*, 22, 577-594.

Ward, J. H. Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.

Wong, V., & Saunders, J. (1993). Business orientation and corporate success. *Journal of Strategic Marketing*, 1, 20-40.

Appendix A: List of Industries

Industry Types	
Agriculture, Forestry and Fishing	Transport, and Storage
Mining	Communication Services
Manufacturing;	Finance and Insurance
Electricity, Gas, and Water Supply	Property and Business Services
Construction	Education
Wholesale Trade	Health and Community Services
Retail Trade	Cultural and Recreational Services
Accommodation, Cafes and Restaurants	Personal and Other Services

Source: (ABS, 2007)

Examining Growth with Statistical Shape Analysis and Comparison of Growth Models

Deniz Sigirli İlker Ercan
Uludag University
Gorukle/Bursa, Turkey

Growth curves have been widely used in the fields of biology, zoology and medicine for assessing some measurable trait of an organism, such as height, weight, area or volume. In statistical shape analysis, a size measure is obtained using the geometrical information of an object as opposed to linear measurements. The performances of commonly used non-linear growth curves are compared by using centroid size as a size measure in a simulation study. An example is provided on the relationship between centroid size of the cerebellum and disease duration in multiple sclerosis patients.

Key words: Centroid size, growth models, statistical shape analysis.

Introduction

Many studies in the field of medicine are related to the examination of geometrical properties of an organ or organism. Although the data sets used in the statistical analyses of medical studies mainly consist of quantitative or qualitative measurement values, an organ or organism's appearance or shape is also used as input data via imaging techniques (Ercan, et al., 2012). Studies performed in medicine and biology commonly evaluate how the shape of an organ or organism is affected by a disease, how the shape is related to covariates such as sex, age or environmental conditions, the comparison of shapes, how to discriminate and classify using shape data, how to describe shape variability, how shape changes during growth and how shape is related to size (Dryden & Mardia, 1998).

Growth patterns can be defined as the composite of geometric changes in biological structure occurring through ontogenetic time (Lele & Richtsmeier, 2001). These changes can be analyzed with growth curve models. The shapes of the growth curves show differences according to the organism type, environmental conditions and the nature of the trait being measured (Colak, et al., 2006). Growth curves seek a model with a biological basis and biologically interpretable parameters (Seber & Wild, 2003).

Several authors have conducted studies in the areas of biology, medicine, zoology and agriculture by assessing some measurable trait of an organism, such as height, weight, area or volume (Carlson & Baremore, 2005; Ersoy, et al., 2007; Karadavut, et al., 2010). In statistical shape analysis, the size measure is obtained by using the geometrical information of an object or an organism, as opposed to considering linear distances or measurements. One of the most commonly used size measures is centroid size. An important feature of centroid size is that it is statistically independent from the shape of the object; this is the only geometrical information that remains when location, scale and rotational effects are filtered out from an object (Dryden & Mardia, 1998). This independence is not valid for other size measures, such as height, weight, area, volume, ratios or angles.

Deniz Sigirli is a Research Assistant in the Department of Biostatistics, Faculty of Medicine. Her research interests are statistical shape analysis and growth curves. Email her at: denizsigirli@hotmail.com. İlker Ercan is a Professor in the Department of Biostatistics, Faculty of Medicine. His research interests are statistical shape analysis, reliability and validity analyses. Email him at: ercan@uludag.edu.tr.

This study compares the performance of commonly used non-parametric growth curve models and examines their efficiency by sample size for each model using centroid size as a size measure. A practical example is given for examining the relationship between centroid size of the cerebellum and the duration of multiple sclerosis (MS) disease in MS patients with the three- and four- parameter logistic, Gompertz and Richards models.

Methodology

Growth Models

Some measure of the size of an object or living thing against time can be modeled using growth curves. In growth studies, both longitudinal and cross-sectional data can be used. Longitudinal data involve responses over time that can be modeled as a stochastic process (Lindsey, 1997). Cross sectional data consist of a group of measures for each age, but each individual is measured only once so that the sample for an age group does not contain any of the individuals in the previous age group. Longitudinal data are particularly useful in studying secular trends and are a requirement for predictive models of development. Alternatively, closely spaced longitudinal data may obscure more general patterns and reveal seemingly erratic, idiosyncratic patterns of individual growth. To study general population patterns, cross-sectional data may be more useful (Lele & Richtsmeier, 2001).

A growth profile will generally be a nonlinear function of time, often reaching an asymptote. In this situation, linear models may not provide adequate explanations for growth; for these types of data, nonlinear models can provide better predictions. Different algorithms are used in nonlinear regression analysis, such as, the Levenberg-Marquardt, the Gauss-Newton and the Newton-Raphson algorithms (Bates & Watts, 1988; Hintze, 2007). A general nonlinear regression model is:

$$Y_i = f(X_i; \theta) + \varepsilon_i \quad (1)$$

$$i = 1, \dots, n.$$

Such a model is reasonable to use with cross-sectional data in which a single size

measurement is obtained for each individual or experimental unit. In equation (1), θ is the predicted parameter's vector and ε_i is the independent and identically distributed error term with mean 0 and variance σ^2 . The X, or independent variable, corresponds to age or another time variable, and Y, the dependent variable, represents the related size measure. Parameters used in growth models – α , β , κ , γ and δ – have biological meanings. The α parameter represents the final size, this parameter also mathematically corresponds to the maximum asymptote point of the curve; β is the initial size and corresponds to the minimum asymptote point of the curve; κ is the parameter that shows the growth rate; γ is the inflection point of the curve; and δ is the second inflection point, which is found in the Richards growth model (Seber & Wild, 2003; Hintze, 2007).

Three Parameter Logistic Model

The three parameter logistic model is an S shaped function:

$$f(x) = \frac{\alpha}{1 + \beta e^{-\kappa x}} \quad (2)$$

$$-\infty < x < \infty.$$

The curve has two asymptotes, when $x \rightarrow -\infty$ as $f(x) = 0$ and when $x \rightarrow \infty$ as $f(x) = \alpha$. Growth typically begins prior to observation when $f(x) > 0$, this can create some difficulties. When $f(x) = \alpha/2$, $x = \gamma$ is obtained and the growth rate reaches a maximum level.

Four Parameter Logistic Model

The four parameter logistic model is an extended form of the three parameter logistic model:

$$f(x) = \gamma + \frac{\alpha - \gamma}{1 + \beta e^{-\kappa x}}. \quad (3)$$

The four parameter model is frequently used in bioassays or immunoassays, such as ELISAs or

dose-response curves (Plikaytis & Carlone, 2005; Wang, et al., 2008; Healy, 1972). In this logistic model, a monotonic function is either always increasing or decreasing for all values of x .

Gompertz Model

The Gompertz model (Gompertz, 1825) was introduced to describe mortality rates in humans (Walter & Bailer, 2005). According to Winsor (1932), Wright first suggested the use of the Gompertz curve for biological growth in 1926. The Gompertz growth curve is given by:

$$f(x) = \alpha e^{\{-e^{-\kappa(x-\gamma)}\}} \quad (4)$$

Richards Model

Developed by Richards in 1959 as a generalization of the classical growth curves, the Richards model is a widely used and flexible growth model with four parameters. This model provides a flexible curve with an arbitrarily placed point of inflection.

$$f(x) = \alpha [1 + (\delta - 1)e^{-\kappa(x-\gamma)}]^{-\frac{1}{1-\delta}}, \quad (5)$$

$\delta \neq 1$.

Other growth functions can be obtained from the Richards function according to the values that δ can take. When $\delta = 0$, a monomolecular growth function is obtained; when $\delta = 2$, a three-parameter logistic function is obtained; when $\delta = 2/3$, and when $\delta \rightarrow 1$, a Gompertz growth function is obtained (Seber & Wild, 2003; Hintze, 2007). The Gompertz, logistic and Richard's growth models have points of inflection and are sigmoid. These models are suitable for quantifying a growth phenomenon that exhibits a sigmoid pattern over time.

Centroid Size

If $X_{k \times m}$ is a $k \times m$ configuration matrix (Cartesian coordinates of k landmarks in m real dimensions) of an object with k landmarks in m dimensions, then a size measure, $g(X)$, is any positive, real value function of the configuration matrix, such that

$$g(aX) = a g(X) \quad (6)$$

for any positive scalar, a . The main size measures used in statistical shape analysis are centroid size, baseline size (as proposed by Galton) and the radius of the inscribed circle for the triangles (as proposed by Miles) (Dryden & Mardia, 1998).

Centroid size is the most frequently used size measure in statistical shape analysis and is:

$$S(X) = \|CX\| = \sqrt{\sum_{i=1}^k \sum_{j=1}^m (X_{ij} - \bar{X}_j)^2}, \quad (7)$$

$X \in R^{km}$

where $\bar{X}_j = \frac{1}{k} \sum_{i=1}^k X_{ij}$ and $\|X\| = \sqrt{\text{trace}(X'X)}$ are the Euclidean norm; C is the centering matrix and is given by

$$C = I_k - \frac{1}{k} 1_k 1_k' \quad (8)$$

where I_k is the $k \times k$ identity matrix and 1_k is the $k \times 1$ vector of ones. Centroid size additionally can be identified as the square root of the sum of the variances of the landmarks around the centroid in x - and y -directions as shown by:

$$S(X) = \sqrt{\sum_{i=1}^k \|(X)_i - \bar{X}\|^2}, \quad (9)$$

where $(X)_i$ is the i^{th} row of the X matrix and $\bar{X} = (\bar{X}_1, \dots, \bar{X}_m)$ is the centroid (Dryden & Mardia, 1998).

Simulation

The original data set, which was used as the reference in the simulation study, consisted of 15 healthy individuals (4 male, 11 female). Corpus callosum (CC) images of these healthy individuals were obtained from the mid-sagittal sections of the magnetic resonance imaging (MRI) scans. The selected landmarks were marked on the digital images using TPSDIG

2.16 software (Rohlf, 2010). The mean, variance and other parameters used for data generation in the simulation study were obtained from this landmark coordinate data set.

The corpus callosum images of the individuals were divided into seven regions according to Witelson's sub-division framework (Witelson, 1989). For the growth curve models, 5th and the 6th regions were combined and analyzed together as one region. A total of 5 landmarks were marked for that region (landmarks 1, 2, 4, 5 and 6). The first 4 landmarks (1, 2, 3 and 4) are the anatomical landmarks defined as in Ozdemir, et al. (2007). To better describe the shape of the brain structure, two additional landmarks (5 and 6) were constructed by referencing these anatomical landmarks. The third landmark was not included in the study but was used in the determination of landmarks 5 and 6 (see Figure 1); a descriptive list of the landmarks is provided in Table 1.

Using age as an independent variable and centroid size as the dependent variable, different growth models were constructed using the original data set. These models are:

1. The three parameter logistic model:

$$y = \frac{(11838.440)}{1 + (85.084)e^{-(1.124)x}} + e$$

2. The four parameter logistic model:

$$y = (362757.700) + \frac{(92,119 - 362757.700)}{1 + (1.788)e^{-(0.427)x}} + e$$

3. The Gompertz model:

$$y = (8953.636)e^{\left(-e^{-(0.002)(x - (-584.773))}\right)} + e$$

4. The Richards model:

$$y = 304.851[1 + (14.756 - 1)e^{-(0.153)(x - (-54.227))}]^{\frac{1}{(1 - 14.756)}} + e$$

In the simulation study, the x values (age) were generated from a normal distribution

Figure 1: The Sub-Divisions of the Corpus Callosum Based on the Witelson Framework and the Landmarks Marked on the 5th and 6th Regions

Table 1: A Descriptive List of the Landmarks Used For the Corpus Callosum (CC)

Landmark Number	Landmark Definition	Landmark Illustration
1	CC-fornix junction	
2	Interior notch of the splenium	
3	Posterior-most point of the CC	
4	Top most point of the CC	
5	The point at which the line that passes through landmark 4 is perpendicular to the segment, which was drawn from landmark 3 to landmark 4, and cuts the lower bound of the CC	
6	The point at which the line that passes through the midpoint of the segment, which was drawn from landmark 3 to landmark 4, is perpendicular to this segment and cuts the upper bound of the CC	

by using the mean and the variance of the age values from the original data set for each model, and error terms, $e \sim N(0, 1)$, were generated. The dependent variable's values (centroid size) were generated using values from the models obtained from the original data set. Simulations were performed for sample sizes $n = 20, 50$ and 100 with 250 repetitions.

Results

To compare the performance of the growth models, mean square error (MSE) criteria were used as given in equation (5). (See Table 2 and Figures 2-3 for MSE values.) The MSE is:

$$MSE = \frac{1}{t} \sum_{j=1}^t \left(\frac{\sum_{i=1}^n (y_{ij} - \hat{y}_{ij})^2}{n - p} \right) \quad (10)$$

where t is the number of replications, p is the number of parameters in the model and n is the sample size in each repetition.

To investigate the efficiencies of the parameter estimates according to sample size, the mean absolute deviations (MAD) and the bias of the estimated coefficients criteria were used. These two criteria were used only to examine each model's performance in itself according to change in the sample size.

To show the difference between the predicted and the actual values of the parameters, the MAD criteria were calculated as:

$$MAD = \frac{1}{t} \frac{1}{p} \sum_{i=1}^t \sum_{j=1}^p |\hat{\beta}_{ij} - \beta_j| \quad (11)$$

where t is the number of replications, p is the number of parameters in the model, n is the sample size in each repetition, $\hat{\beta}_{ij}$ is the predicted value of the j^{th} parameter in the i^{th} model and β_j is the actual value of the j^{th} parameter. (See Table 3 and Figure 4.)

STATISTICAL SHAPE ANALYSIS AND COMPARISON OF GROWTH MODELS

Table 2: MSE Values for Growth Models

n	Three Parameter Logistic	Four Parameter Logistic	Gompertz	Richards
20	1.06808	1.0752	1.05095	1.42999
50	1.01298	1.01018	1.00929	1.03037
100	0.95822	0.96118	0.95646	0.96734

Figure 2: MSE Values for Growth Models

Figure 3: Percentage Change for MSE Values

Table 3: MAD Values for Growth Models

n	Three Parameter Logistic	Four Parameter Logistic	Gompertz	Richards
20	3061.241	5438899	2076.513	374.8195
50	3007.479	372227.4	1954.277	330.3648
100	2986.007	299968	1997.179	96.05916

Figure 4: Percentage Changes for MAD Values

To show the difference between the mean of the predicted values and the actual values of the parameters, the bias criteria were calculated using

$$\text{Bias} = \frac{1}{p} \sum_{j=1}^p \left| \frac{\sum_{i=1}^t \hat{\beta}_{ij}}{t} - \beta_j \right| \quad (12)$$

where t is the number of replications, p is the number of parameters in the model, n is the sample size in each repetition, $\hat{\beta}_{ij}$ is the predicted value of the j^{th} parameter in the i^{th}

model and β_j is the actual value of the j^{th} parameter. (See Table 4 and Figure 5.)

Practical Example

The data set used in this example consists of the MRI scans of 44 (17 (38.64%) male, 27 female (61.36%)) multiple sclerosis (MS) patients. The mean age was 32.07 ± 8.46 (mean $\pm$ standard deviation) years. The median duration of the MS disease was 25 (4-72) months (median (min-max)). All MS patients fit the McDonald, et al., 2001 criteria. An institutional review board approved the retrospective study and all participants gave informed consent prior to the start of the study.

STATISTICAL SHAPE ANALYSIS AND COMPARISON OF GROWTH MODELS

Table 4: Bias Values for Growth Models

n	Three Parameter Logistic	Four Parameter Logistic	Gompertz	Richards
20	3009.35	5392501	2063.3	339.116
50	3007.48	240113	1954.28	279.614
100	2953.82	157085.5	1996.45	46.7993

Figure 5: Percentage Changes for Bias Values

Eight midline cerebellar landmarks were selected from the image of the mid-sagittal plane (see Figure 6). The landmarks were chosen on the basis of reliability, significant anatomical coverage and previous cerebellar morphological descriptions in MS patients. A descriptive list of these anatomical landmarks is provided in Table 5. The relationship between centroid size of cerebellum and the duration of the MS disease was examined using the three and four parameter logistic, Gompertz and Richards models. The mean squared error and R^2 for the models are shown in Table 6.

Among the models of the studied relationship, the Gompertz model and three parameter logistic model had lower MSE values,

while the four parameter logistic model had the highest R^2 value. All models significantly predicted the relationship between centroid size and the duration of disease. Figures 7-10 show that a decrease occurs in the cerebellum size of the MS patients as the duration of disease increases.

Conclusion

With the technological advances in the fields of biology and medicine, different methods have been developed to analyze an organ's or an organisms' forms by recording the geometrical locations of landmarks. Statistical shape analysis plays an important role in such studies.

Figure 6: T1-Weighted Mid-Sagittal Slice Demonstrating the Cerebellar Landmarks

Table 5: Descriptive List of the Landmarks Used for the Cerebellum

Landmark Number	Landmark Definition
1	Velum medullare superius angulation-cerebellar outline junction
2	Superior cerebellum
3	Primary fissure- cerebellar outline junction
4	Posterior cerebellum
5	Prepyramidal fissure- cerebellar outline junction
6	Inferior cerebellum
7	Velum medullare inferius angulation-cerebellar outline junction
8	Fastigium cerebelli

Table 6: Growth Models of the Relationship between Cerebellum Size and Disease Duration*

Model	Predicted Equation	R^2	MSE	p-value
Three Parameter Logistic	$CS = \frac{(3.89)}{1 + (0.35)e^{-(0.009)DD}} + e$	0.26992	0.04849	0.00158
Four Parameter Logistic	$CS = (3.37) + \frac{(1.57 - 3.37)}{1 + (0.37)e^{-(0.017)DD}} + e$	0.26993	0.04970	0.00524
Gompertz	$CS = (159.47)e^{\left(-e^{-(0.006)(DD-(4.27))}\right)} + e$	0.26992	0.04849	0.00158
Richards	$CS = 418.03[1 + (2.81 - 1)e^{-(1.39)(DD-(19.89))}]^{\frac{1}{1-2.81}} + e$	0.27445	0.04940	0.00468

*CS: centroid size; DD: disease duration; R^2 : the coefficient of determination; MSE: Mean Squared Error

Figure 7: Three Parameter Logistic Model

Figure 8: Four Parameter Logistic Model

Figure 9: Gompertz Model

Figure 10: Richards Model

Biological processes, such as disease or injury, ontogenetic development or adaptation to local geographic factors can cause shape differences between individuals. These differences in shape may signal differences in the processes of growth and morphogenesis. A shape analysis is one approach to understanding the diverse causes of variation and morphological transformation (Zelditch, 2004).

Growth studies produce important information on aspects of the biology of an organism, such as the genetic basis of morphogenesis, the phylogenetic underpinnings of developmental patterns or the role of hormones, teratogens, dietary elements and other environmental variables on the growth processes (Lele & Richtsmeier, 2001). The relationship between a measurable trait of an organism and time can be modeled with a growth curve. Several applied studies have been performed using growth curves and taking a measurable trait such as area, length, weight or volume as a size measurement (Carlson & Baremore, 2005; Ersoy, et al., 2007; Karadavut, et al., 2010; Ozel & Ertekin, 2001); however, in a statistical shape analysis, size measurement is obtained by using geometrical information about an object or organism. One commonly used size measure in shape analysis is centroid size (Dryden & Mardia, 1998). An important feature of centroid size is that it is independent from the shape; this feature is not valid for the other size measures, such as length, weight, area or volume.

This study used centroid size, as opposed to the classical measurements used in nonlinear growth curves. In the literature, especially in the field of geometric morphometry, several studies have investigated the relationship between size and age. In these studies, linear models have usually been used with the natural logarithm of the centroid size as the dependent variable and age as the independent variable. The only study in the literature where non-linear growth models were used with centroid size was the size measure study by Colak, et al. (2011). This illustrates the necessity of investigating the performance of non-linear growth models where centroid size is the dependent variable. Cardini and Elton (2007) investigated the effect of sample size on geometric morphometric studies of size and

shape; they note that sampling error might affect estimates of the statistical parameters – this observation was virtually absent in geometric morphometrics and few studies have performed simulations and mathematical modeling to theoretically examine the issues (Cardini & Elton, 2007). It appears that this is the first study to compare non-linear growth models according to sample size by using centroid size as a size measure.

Summary

In all growth models examined in this study, the MSE decreased as the sample size increased. The Richards model had the largest MSE values in small sample sizes of all the models. As the sample size increased, the MSE value of the Richards model become lower, reaching a comparable value to the MSE values of the other models. Therefore, the Richards model is not suitable for small sample sizes. The Gompertz model and the three and four parameter logistic models had similar MSE values for all sample sizes and experienced similar effects from the decrease of sample size. Except in the small sample size condition, there were no major differences between the models in terms of MSE values.

When the growth curves were assessed in terms of the MAD measure, there was a decrease in the MAD values of the Richards model and the three parameter and four parameter logistic models; however, there was a slight increase in the MAD value of the Gompertz model as sample size increased. The three parameter logistic and Gompertz models showed the lowest decrease in MAD as sample size increased. The four parameter logistic model experienced the largest effects from changes in sample size and exhibited the largest percent change decrease in its MAD value. While transitioning from a moderate to a large sample size, the Richards model showed a significant decrease in MAD value, but the Richards model did not show a remarkable decrease in transitioning from a small to a moderate sample size.

Results for the bias measure were similar to the results for the MAD measure. Although a decrease was observed in the bias values of the Richards model and the three and

four parameter logistic models, there was a small increase in the bias value of the Gompertz model in transitioning to a large sample size. The models that showed the smallest decrease in bias with the increase in the sample size were the three parameter logistic model and the Gompertz model. The four parameter logistic model was the model most affected by sample size, and it was the model that had the largest decrease in its bias value. When transitioning from a moderate to a large sample size the Richards model showed a large decrease in bias but it did not show a remarkable decrease in transitioning from a small to a moderate sample size.

Generally, the Richards model is not convenient for small samples in terms of both model performance and parameter estimates. The three parameter logistic and Gompertz models do not display differences in parameter estimates by sample size, therefore, the three parameter logistic and Gompertz models are preferable to the other two models, particularly for small samples.

References

- Bates, D. M., Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York, NY: John Wiley & Sons.
- Cardini, A., & Elton, S. (2007). Sample size and sampling error in geometric morphometric studies of size and shape. *Zoomorphology*, 126, 121-134.
- Carlson, J., & Baremore, I. (2005). Growth dynamics of the spinner shark (*Carcharhinus brevipinna*) off the United States southeast and Gulf of Mexico coasts: A comparison of methods. *Fishery Bulletin*, 103, 280-291.
- Colak, C., Orman, M. N., & Ertugrul, O. (2006). Simple linear and logistic growth model for the body measurements of Simmental x Southern Anatolian Red crossbred cattle. *Veterinary Journal of Ankara University*, 3, 195-199.
- Colak, C., Ercan, I., Dogan, M., Ozdemir, S. T., Sener, S., & Alkan, A. (2011). Detecting the shape differences of the corpus callosum in Behçet's disease by statistical shape analysis. *Anatomical Records*, 294, 870-874.
- Dryden, I. L., & Mardia, K. V. (1998). *Statistical shape analysis*. New York, NY: Wiley.
- Ercan, I., Ocakoglu, G., Sigirli, D., & Ozkaya, G. (2012). Statistical shape analysis and usage in medical sciences. *Turkiye Klinikleri Journal of Biostatistics*, 4(1), 27-35.
- Ersoy, L., Mendes, M., & Keskin, S. (2007). Estimation of parameters of linear and nonlinear growth curve models at early stage in California turkeys. *Archiv fur Geflugelkunde*, 71, 175-180.
- Healy, M. J. R. (1972). Statistical analysis of radioimmunoassay data. *Biochemical Journal*, 130, 207.
- Hintze, J. (2007). *NCSS and PASS: Number cruncher statistical systems*. Utah: NCSS Statistical & Power Analysis Software Publications.
- Karadavut, Palta, U. C., Kokten, K., & Bakoglu, A. (2010). Comparative study on some non-linear growth models for describing leaf growth of maize. *International Journal of Agricultural Biology*, 12, 227-230.
- Lele, S., Richtsmeier, J. T. (2001). *An invariant approach to statistical analysis of shapes*. Florida: Chapman & Hall/CRC Press.
- Lindsey, J. K. (1997). *Applying generalized linear models*. New York, NY: Springer Verlag.
- McDonald, W. I., Compston, A., Edan, G., Goodkin, D., Hartung, H.P., Lublin, F.D., McFarland, H. F., Paty, D. W., Polman, C. H., Reingold, S. C., Sandberg-Wollheim, M., Sibley, W., Thompson, A., Van Den Noort, S., Weinshenker, B. Y., Wolinsky, J. S. (2001). Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Annals of Neurology*, 50(1), 121-127.

Ozdemir, S. T., Ercan, I., Sevinc, O., Guney, I., Ocaoglu, G., Aslan, E., & Barut, C. (2007). Statistical shape analysis of differences in the shape of the corpus callosum between genders. *Anatomical Records*, 290, 825-830.

Ozel, H. B., & Ertekin, M. (2011). Growth models in investigating oriental beech (*Fagus orientalis* Lipsky.) juvenilities growth performance in the Western Black Sea in Turkey (Devrek-Akçasu Case Study). *Romanian Biotechnological Letters*, 16, 5851.

Plikaytis, B.D., Carlone, G.M. (2005). Statistical considerations for vaccine immunogenicity trials. Part 1: introduction and bioassay design and analysis. *Vaccine*. 23: 1596-1605.

Rohlf, F. J. (2010). *TPSDIG Version 2.16*. New York: Ecology and Evaluation, SUNY at Stony Brook.

Seber, G. A. F., & Wild, C. J. (2003). *Nonlinear regression*. Hoboken, NJ: John Wiley & Sons.

Walter, W. P., & Bailer, A. J. (2005). Analyzing environmental data. *Statistical Papers*, 51, 1009-1010.

Wang, D., Burton, R. L., Nahm, M. H., & Soong, S. J. (2008). A four-parameter logistic model for estimating titers of functional multiplexed pneumococcal opsonophagocytic killing assay. *Journal of Biopharmaceutical Statistics*, 18, 307-325.

Winsor, C. P. (1932). The Gompertz curve as a growth curve. *Proceedings of the National Academy of Sciences*, 18(1), 1-8.

Witelson, S. F. (1989). Hand and sex differences in the isthmus and genu of the human corpus callosum. *Brain*, 112, 799.

Zelditch, M. (2004). *Geometric morphometrics for biologists: A primer*. New York, NY: Elsevier Academic Press.

Extreme Value Charts and Analysis of Means (ANOM) Based on the Log Logistic Distribution

B. Srinivasa Rao
R. V. R. & J. C. College of
Engineering,
Guntur, Andhrapradesh, India

J. Pratapa Reddy
St. Ann's College for Women,
Guntur, Andhrapradesh, India

G. Sarath Babu
Chebrolu Hanumaiah Institute of
Pharmaceutical Sciences,
Guntur, Andhrapradesh, India

A probability model of a quality characteristic is assumed to follow a log logistic distribution. This article proposes variable control charts, termed extreme value charts, based on the extreme values of each subgroup. The control chart constants depend on the probability model of the extreme order statistics and the size of each subgroup. The analysis of means (ANOM) technique for a skewed population is applied with respect to log logistic distribution. Results are illustrated using examples based on real data.

Key words: ANOM, LLD, in control, equi-tailed, Q-Q plot.

Introduction

The probability density function (PDF) of a log logistic distribution (LLD) with shape parameter b and scale parameter σ is given by

$$f(x, b, \sigma) = \frac{b \left(\frac{x}{\sigma} \right)^{b-1}}{\sigma \left(1 + \left(\frac{x}{\sigma} \right)^b \right)^2},$$

$$x > 0, \sigma > 0, b > 1 \quad (1.0.1)$$

and its cumulative distribution function (CDF) is

$$F(x, b, \sigma) = \frac{\left(\frac{x}{\sigma} \right)^b}{1 + \left(\frac{x}{\sigma} \right)^b},$$

$$x > 0, \sigma > 0, b > 1. \quad (1.0.2)$$

When $\sigma = 1$ and $b > 1$ these equations are termed standard PDF and CDF. In order to construct a control chart using extreme observations of a subgroup drawn from a production process with the quality variate following a LLD, the percentiles of extreme order statistics from LLD samples are needed. Specifically, the test statistic on the extreme value control chart is the original sample vector $X = (x_1, x_2, \dots, x_n)$ from ongoing production. In this chart all individual sample observations are plotted into the control chart without calculating any statistics. A corrective action is taken after one, or either, of the extreme values – namely $x_{(1)}$ (sample minimum) and $x_{(n)}$ (sample maximum) – of the sample respectively fall above or below specified lines (limits); this is why the chart is called an extreme value controlled chart.

The Shewart (1986) controlled chart is a common quality control statistical tool: When a Shewart chart indicate the presence of an assignable cause, a process adjustment can be made if the remedy is known; otherwise the

B. Srinivasa Rao is an Associate Professor of Statistics in the Department of Mathematics and Humanities. Email him at: boyapatisrinu@yahoo.com. J. Pratapa Reddy is an Associate Professor of Statistics in the Department of Computer Applications. Email him at: jakkampratapa@yahoo.co.in. G. Sarath Babu is an Assistant Professor of Statistics. Email him at: gorantla.sarath@gmail.com.

suspected presence of assignable cause is regarded as an indication of heterogeneity of the subgroup statistic for which the control chart was developed. For example, if the statistic is the sample mean, this leads to heterogeneity of the process mean and indicates departures from the target mean. Such an analysis is generally carried out by dividing a collection of a given number of subgroup means into categories, such that means within a category are homogenous and those between categories are heterogeneous. This procedure, developed by Ott (1967) is called analysis of means (ANOM).

When using the ANOM technique the concept of a control chart for means is viewed differently, grouping of plotted means that fall within or outside control limits. For the homogeneity of the means it is necessary that all means fall within the control limits. If $(1 - \alpha)$ is the confidence coefficient, then the probability that all subgroup means will fall within the control limits is $(1 - \alpha)$. Assuming independence of the subgroup, the probability statement becomes n^{th} power of the probability that a subgroup mean will fall within the limits. This means that, in the sampling distribution of $\bar{x}$, the confidence interval for $\bar{x}$ to lie between two specified limits should be equal to $(1 - \alpha)^{1/n}$. This same principle is adapted through log logistic distribution in this study. This article explores ANOM using control limits of extreme value statistics considering only control chart aspects. (See Rao (2005) for a detailed description of ANOM; other related works include: Ramig, 1983; Bakir, 1994; Bernard & Wludyka, 2001; Wludyka, et al., 2001; Montgomery, 2001; Nelson & Dudewicz, 2002; Rao & Prankumar, 2002; Farnum, 2004; Guirguis & Tobias, 2004; Srinivasa Rao & Kantam, 2012.)

Extreme Value Charts

The given sample observations are assumed to follow log logistic mode. The controlled lines are determined by the theory of extreme order statistics based on a half logistic model. The controlled lines are determined in such a way that an arbitrarily chosen x_i of $X = (x_1, x_2, \dots, x_n)$ lies with the probability $(1 - \alpha)^{1/n}$ within the limits. This can be formulated as a

probability inequality as: $P(x_i \leq L) = \alpha/2$ and $P(x_n \geq U) = \alpha/2$. The theory of order statistics states that the cumulative distribution function of the least and highest order statistics in a sample of size n from any continuous population are $[F(x)]^n$ and $1 - [1 - F(x)]^n$, respectively, where $F(x)$ is a cumulative distribution function (CDF) of the population. If $1 - \alpha$ is desired at 0.9973, then α would equal 0.0027. Taking $F(x)$ as the CDF of a standard log logistic model results in solutions of the equations $1 - [1 - F(x)]^n = 0.00135$ and $[F(x)]^n = 0.99865$ which, in turn, can be used to develop the controlled limits of an extreme value chart. The solutions for the two equations for $n = 2$ (1) 10 with $b = 2, 3, 4$ and 5 are shown in table 2.1 and denoted as $Z^* = Z_{(1)0.00135}$ and $Z^{**} = Z_{(n)0.99865}$.

The values shown in table 2.1 indicate the following probability statements:

$$P\left(\begin{matrix} Z_{(1)0.00135} < Z_i < Z_{(n)0.99865}, \\ \forall i, = 1, 2, \dots, n \end{matrix}\right) = 0.9973 \quad (2.0.3)$$

and

$$P\left(\begin{matrix} \sigma Z_{(1)0.00135} < x_i < \sigma Z_{(n)0.99865}, \\ \forall i, = 1, 2, \dots, n \end{matrix}\right) = 0.9973 \quad (2.0.4)$$

Taking $\frac{\bar{x}}{1.5708}$, $\frac{\bar{x}}{1.0472}$, $\frac{\bar{x}}{0.7854}$ and $\frac{\bar{x}}{0.6283}$ as an unbiased estimates of σ when $b = 2$, $b = 3$, $b = 4$ and $b = 5$, respectively, the equation becomes

$$P\left(\begin{matrix} D_3^* \bar{x} < x_i < D_4^* Z_{(n)0.99865}, \\ \forall i, = 1, 2, \dots, n \end{matrix}\right) = 0.9973 \quad (2.0.5)$$

where, for $b = 2$:

$$D_3^* = \frac{Z(1)(0.00135)}{1.5708}$$

and

$$D_4^* = \frac{Z(n)(0.99865)}{1.5708}.$$

and

$$D_4^* = \frac{Z(n)(0.99865)}{0.7854}.$$

For b = 3:

$$D_3^* = \frac{Z(1)(0.00135)}{1.0472}$$

For b = 5:

$$D_3^* = \frac{Z(1)(0.00135)}{0.6283}$$

and

$$D_4^* = \frac{Z(n)(0.99865)}{1.0472}.$$

and

$$D_4^* = \frac{Z(n)(0.99865)}{0.6283}.$$

For b = 4:

$$D_3^* = \frac{Z(1)(0.00135)}{0.7854}$$

Thus, D_3^* and D_4^* constitute the control chart constants for the extreme value charts (see Table 2.2 for n = 2(1)10).

Table 2.1: Control Chart Limits of Extreme Value Charts

n	b=2		b=3		b=4		b=5	
	Z_*	Z_{**}	Z_*	Z_{**}	Z_*	Z_{**}	Z_*	Z_{**}
2	0.0259	38.4705	0.0877	11.3959	0.1612	6.2024	0.2322	4.3057
3	0.0212	47.1192	0.0766	13.0456	0.1456	6.8643	0.2141	4.6695
4	0.0183	54.4101	0.0696	14.3588	0.1355	7.3763	0.2021	4.9462
5	0.0164	60.8334	0.0646	15.4677	0.1282	7.7995	0.1933	5.1719
6	0.0150	66.6404	0.0608	16.4371	0.1225	8.1633	0.1864	5.3640
7	0.0138	71.9804	0.0577	17.3038	0.1178	8.4841	0.1807	5.5320
8	0.0129	76.9508	0.0552	18.0915	0.1139	8.7721	0.1760	5.6817
9	0.0122	81.6190	0.0531	18.8160	0.1106	9.0343	0.1719	5.8172
10	0.0116	86.0343	0.0513	19.4886	0.1078	9.2754	0.1683	5.9411

Analysis of Means (ANOM): Log Logistic Distribution

When the data variate follows log logistic distribution, suppose $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ are arithmetic means of k subgroups of size n drawn from a log logistic model. The subgroups means are used to develop control charts to assess whether the population from which these subgroups are drawn is operating with admissible quality variations. Depending on the basic population model, control chart constants may be used. In general, the process may be said to be in control if all subgroup means are within the control limits; otherwise the process is said to lack control. If α is the level of significance of this decision, the following probability statements apply:

$$P(LCL < x_i, \forall i, = 1, 2, \dots, k < UCL) = 1 - \alpha \quad (3.0.6)$$

using the notion of independent subgroups,

(3.0.6) becomes

$$P(LCL < x_i < UCL) = (1 - \alpha)^{1/k} \quad (3.0.7)$$

With equi-tailed probability for each subgroup mean, two constants, for example L^* and U^* , may be found such that

$$P(x_i < L^*) = P(\bar{x}_i > U^*) = \frac{1 - (1 - \alpha)^{1/k}}{2} \quad (3.0.8)$$

For skewed populations, such as the LLD, it is necessary to calculate L^* , U^* separately from the sampling distribution of $\bar{x}_i$. Accordingly, these depend on the subgroup size n and number of subgroups k . The percentiles of the sampling distribution of $\bar{x}$ in samples from a log logistic distribution for $b = 2, b = 3, b = 4$ and $b = 5$ with $\sigma = 1$ were calculated using Monte-Carlo simulations (see Tables 3.1, 3.2, 3.3 and 3.4).

Table 2.2: Control Chart Limits of Extreme Value Charts

n	b = 2		b = 3		b = 4		b = 5	
	D_3^*	D_4^*	D_3^*	D_4^*	D_3^*	D_4^*	D_3^*	D_4^*
2	0.0165	24.4910	0.0837	10.8823	0.2052	7.8971	0.3695	6.8529
3	0.0135	29.9969	0.0731	12.4576	0.1854	8.7398	0.3423	7.4319
4	0.0116	34.6384	0.0664	13.7116	0.1726	9.3917	0.3216	7.8723
5	0.0104	38.7276	0.0617	14.7705	0.1632	9.9306	0.3076	8.2315
6	0.0095	42.4246	0.0580	15.6962	0.1559	10.3938	0.2966	8.5373
7	0.0088	45.8240	0.0551	16.5238	0.1500	10.8022	0.2876	8.8047
8	0.0082	48.9882	0.0527	17.2760	0.1450	11.1690	0.2801	9.0429
9	0.0078	51.9601	0.0507	17.9679	0.1408	11.5028	0.2735	9.2586
10	0.0074	54.7710	0.0489	18.6102	0.1372	11.8098	0.2678	9.4558

Table 3.1: Percentiles of Sample Mean in LLD with $b = 2$

n	0.99865	0.99	0.975	0.95	0.05	0.025	0.01	0.00135
2	12.9560	4.8648	3.2368	2.363	0.2646	0.2144	0.1624	0.1002
3	11.3630	4.4994	3.0711	2.3002	0.3390	0.2887	0.2368	0.1772
4	10.5971	3.8004	2.7757	2.0875	0.3848	0.3313	0.2767	0.2002
5	9.8124	3.5276	2.5244	1.9696	0.4228	0.3709	0.3180	0.2237
6	9.6424	3.47156	2.4932	1.9646	0.4518	0.4008	0.3446	0.2658
7	7.0472	3.2050	2.3660	1.9001	0.4692	0.4176	0.3662	0.2896
8	6.5408	2.9775	2.2500	1.8469	0.4965	0.4435	0.3914	0.3202
9	5.8094	2.8794	2.2181	1.8242	0.5144	0.4681	0.4201	0.3434
10	6.1672	2.7692	2.2176	1.8063	0.5284	0.4827	0.4331	0.3635

Table 3.2: Percentiles of Sample Mean in LLD with $b = 3$

n	0.99865	0.99	0.975	0.95	0.05	0.025	0.01	0.00135
2	4.8927	3.1605	2.6322	2.2986	0.7921	0.7115	0.6210	0.4927
3	4.3935	2.7927	2.4511	2.1816	0.8913	0.8211	0.7476	0.6406
4	3.3256	2.6008	2.2806	2.0491	0.9423	0.8750	0.8007	0.6739
5	3.1616	2.4128	2.1578	1.9784	0.9857	0.9256	0.8598	0.7290
6	3.2322	2.3737	2.1150	1.9448	1.0203	0.9629	0.8933	0.7799
7	2.9030	2.2476	2.0506	1.9071	1.0350	0.9802	0.9162	0.8115
8	2.7292	2.1760	2.0073	1.8728	1.0608	1.0041	0.9447	0.8522
9	2.5742	2.1361	1.9732	1.8443	1.0779	1.0272	0.9693	0.8833
10	2.5429	2.1235	1.9520	1.8268	1.0961	1.0490	0.9930	0.9017

Table 3.3: Percentiles of Sample Mean in LLD with $b = 4$

n	0.99865	0.99	0.975	0.95	0.05	0.025	0.01	0.00135
2	6.2330	3.4078	2.6332	2.1740	0.5154	0.4459	0.3739	0.2719
3	5.8539	2.9913	2.4142	2.0571	0.6056	0.5447	0.4763	0.3970
4	4.0852	2.6589	1.2396	1.8961	0.6549	0.5949	0.5293	0.4205
5	4.0895	2.4914	2.0797	1.8209	0.6973	0.6402	0.5764	0.4649
6	3.9497	2.4311	2.0150	1.7868	0.7278	0.6728	0.6120	0.5129
7	3.3344	2.2616	1.9481	1.7396	0.7435	0.6919	0.6317	0.5374
8	3.1361	2.1661	1.8742	1.7040	0.7702	0.7172	0.6603	0.5799
9	2.8196	2.0899	1.8578	1.6766	0.7869	0.7402	0.6875	0.6204
10	2.9277	2.1209	1.8353	1.6523	0.8039	0.7567	0.7036	0.6275

Table 3.4: Percentiles of Sample Mean in LLD with $b = 5$

n	0.99865	0.99	0.975	0.95	0.05	0.025	0.01	0.00135
2	4.5322	3.2326	2.8121	2.5223	1.0811	0.9880	0.8909	0.7344
3	4.1151	2.9163	2.6339	2.4091	1.1882	1.1093	1.0315	0.9027
4	3.3010	2.7435	2.4671	2.2876	1.2398	1.1642	1.0868	0.9514
5	3.0996	2.5609	2.3640	2.2206	1.2812	1.2195	1.1485	1.0053
6	2.9154	2.4838	2.3201	2.1834	1.3056	1.2545	1.1927	1.0291
7	2.8736	2.4256	2.2716	2.1508	1.3344	1.2759	1.2103	1.0960
8	2.7778	2.3577	2.2310	2.1206	1.3587	1.3036	1.2420	1.1449
9	2.6438	2.3260	2.1990	2.0882	1.3776	1.3254	1.2603	1.1749
10	2.6299	2.3118	2.1803	2.0749	1.3973	1.3463	1.2899	1.1963

The percentiles shown in Tables 3.1 – 3.4 are used in equation (3.0.8) for specified n and k to determine L^* and U^* for $\alpha = 0.05$ (see Tables 3.5, 3.6, 3.7 and 3.8). A control chart for averages showing in control conclusions indicates that all subgroups means, though varying among themselves, are homogenous in

some cells. This is the null hypothesis in an analysis of variance technique, hence, the constants shown in tables 3.5 - 3.8 can be used as an alternative to analysis of variance techniques. For a normal population Ott's (1967) tables can be used, and for a LLD the tables shown herein can be used.

Table 3.5: LLD Constants for Analysis of Means for $b = 2$, $(1-\alpha) = 0.95$

k	n									
	2		3		4		5		6	
1	0.3303	5.1477	0.4539	4.7591	0.5212	4.3511	0.5768	4.0660	0.6266	3.9113
2	0.2639	7.0248	0.9864	6.6701	0.4599	5.8876	0.5161	5.4263	0.5798	4.9528
3	0.2365	8.9046	0.3569	7.7521	0.4292	6.9287	0.4819	6.2444	0.5517	5.7732
4	0.2192	10.4238	0.3355	8.9742	0.4080	7.7136	0.4591	7.0343	0.5307	6.3341
5	0.2068	11.5178	0.3169	9.5190	0.3742	8.4092	0.4451	7.4928	0.5140	7.3627
6	0.2015	12.4982	0.3054	10.1839	0.3697	9.1787	0.4303	8.0430	0.4153	7.5537
7	0.1972	12.9499	0.3007	10.8149	0.3592	9.9402	0.4199	8.7234	0.4779	7.8955
8	0.1890	13.4191	0.2950	11.0937	0.3544	10.4101	0.4131	8.8267	0.4708	8.3140
9	0.1866	13.9979	0.2931	12.1805	0.3532	10.8970	0.4056	9.1813	0.4665	8.8593
10	0.1716	14.6854	0.2626	12.4166	0.3374	11.0593	0.3878	9.6395	0.4592	9.0119
20	0.1331	17.9113	0.2247	16.9479	0.3130	14.8274	0.3423	12.3236	0.4244	11.9848
30	0.1028	23.7632	0.2097	20.7346	0.3016	18.0296	0.3300	14.8560	0.3885	13.4620
40	0.0946	33.6935	0.1780	24.0424	0.2751	19.7351	0.3118	16.4754	0.3687	16.1631
50	0.0929	34.7754	0.1686	25.1053	0.2745	22.5189	0.3088	18.1307	0.3637	16.4374

k	n							
	7		8		9		10	
1	0.6592	3.6837	0.6945	3.6131	0.7253	3.5156	0.7536	3.4029
2	0.5999	4.6640	0.6423	4.5937	0.6657	4.4740	0.6873	4.2973
3	0.5712	5.4824	0.6103	5.2898	0.6424	5.1286	0.6513	4.8740
4	0.5561	5.9729	0.5949	5.7324	0.6134	5.7535	0.6309	5.0412
5	0.5378	6.7590	0.5838	6.2880	0.6025	6.2461	0.6199	5.8142
6	0.5280	7.1752	0.5682	6.9044	0.5905	6.6190	0.6140	6.0589
7	0.5243	7.7579	0.5596	7.2190	0.5833	6.9793	0.5962	6.3483
8	0.5224	7.9220	0.5499	7.6895	0.5821	7.2779	0.5916	6.6536
9	0.5213	8.6406	0.5493	7.8990	0.5800	7.6447	0.5863	7.0156
10	0.5022	8.6938	0.5437	8.0013	0.5672	7.7415	0.5752	7.0463
20	0.4771	11.6854	0.5106	10.6656	0.5192	9.8981	0.5427	9.9847
30	0.4510	13.1287	0.4994	12.2854	0.5059	11.3192	0.5237	10.3004
40	0.4249	14.5202	0.4815	14.1756	0.4751	13.1701	0.4990	12.7299
50	0.4242	16.1104	0.4777	14.5343	0.4677	13.2737	0.4972	13.4114

EXTREME VALUE CHARTS AND ANOM BASED ON LOG LOGISTIC DISTRIBUTION

Table 3.6: LLD Constants for Analysis of Means for $b = 3$, $(1-\alpha) = 0.95$

k	n									
	2		3		4		5		6	
1	0.4695	2.7584	0.5683	2.5524	0.6248	2.3464	0.6666	2.2172	0.6988	2.1238
2	0.4059	3.3751	0.5101	3.0785	0.5720	2.7479	0.6177	2.5241	0.6660	2.3903
3	0.3733	3.8157	0.4875	3.3458	0.5469	2.9638	0.5929	2.8291	0.6448	2.5548
4	0.3593	4.1241	0.4584	3.6189	0.5203	3.0344	0.5695	2.9484	0.6255	2.6804
5	0.3427	4.5531	0.4451	3.8694	0.5063	3.4598	0.5526	0.0452	0.6095	2.8544
6	0.3388	4.7973	0.4414	3.9541	0.4956	3.5578	0.5437	3.1117	0.5999	2.9205
7	0.3368	4.9647	0.4295	4.0572	0.4881	3.6225	0.5349	3.2163	0.5868	3.0714
8	0.3145	5.2882	0.4167	4.1558	0.4787	3.6740	0.5208	3.3446	0.5772	3.1013
9	0.3136	5.3570	0.4015	4.2656	0.4700	3.7688	0.5152	3.3985	0.5730	3.2142
10	0.3099	5.4002	0.3986	4.3335	0.4645	3.8646	0.5135	3.4171	0.5712	3.2412
20	0.2624	6.1439	0.3477	5.6020	0.4415	4.6598	0.4769	4.1357	0.5375	3.6726
30	0.2455	8.6207	0.3273	6.1971	0.4184	5.0747	0.4486	4.4602	0.5233	4.0896
40	0.2144	8.7280	0.2993	6.3545	0.4108	5.4735	0.4358	4.5999	0.5046	4.3879
50	0.2044	9.0887	0.2835	6.6497	0.3955	5.7045	0.4300	4.6674	0.5240	4.4110

k	n							
	7		8		9		10	
1	0.7261	2.0437	0.7474	1.9820	0.7730	1.9419	0.7902	1.9098
2	0.6783	2.2991	0.7126	2.2268	0.7322	2.1810	0.7442	2.1333
3	0.6559	2.4746	0.6947	2.3405	0.7071	2.3472	0.7205	2.2488
4	0.6404	2.5873	0.6761	2.4314	0.6957	2.4292	0.7071	2.3594
5	0.6281	2.6892	0.6663	2.8787	0.6826	2.5798	0.6962	2.4315
6	0.6225	2.8124	0.6627	2.6851	0.6744	2.6859	0.6848	2.5297
7	0.6145	2.8789	0.6419	2.7863	0.6701	2.7588	0.6824	2.5859
8	0.6110	3.0330	0.6362	2.8695	0.6638	2.7963	0.6731	2.6091
9	0.6087	3.1342	0.6344	2.9247	0.6527	2.8245	0.6703	2.6647
10	0.6048	3.1402	0.6317	2.9568	0.6522	2.8585	0.6652	2.7014
20	0.5782	3.6224	0.6163	3.2793	0.6165	3.1551	0.6323	2.9422
30	0.5664	3.9450	0.6036	3.6797	0.6092	3.6316	0.6161	3.3713
40	0.5332	4.0987	0.5951	3.8300	0.5934	3.5372	0.6112	3.3857
50	0.5277	4.2365	0.5902	3.9501	0.5824	3.6456	0.6035	3.4009

Table 3.7: LLD Constants for Analysis of Means for $b = 4$, $(1-\alpha) = 0.95$

k	n									
	2		3		4		5		6	
1	0.5538	2.0708	0.6462	1.9460	0.6916	1.8023	0.7255	1.7171	0.7522	1.6570
2	0.4998	2.3840	0.5967	2.1913	0.6472	1.9817	0.6821	1.8793	0.7205	1.8782
3	0.4692	2.5861	0.5711	2.3067	0.6235	2.0966	0.6641	1.9984	0.7016	1.8626
4	0.4501	2.7621	0.5495	2.4355	0.6032	2.2188	0.6423	2.0633	0.6858	1.9325
5	0.4424	2.9285	0.5382	2.5761	0.5927	2.3001	0.6290	2.1164	0.6741	1.9779
6	0.4363	2.9917	0.5261	2.6173	0.5881	2.3841	0.6182	2.1564	0.6696	2.0239
7	0.4280	3.0573	0.5194	2.6437	0.5769	2.4234	0.6101	2.1760	0.6782	2.0083
8	0.4122	3.1024	0.5090	2.6733	0.5636	2.4560	0.6063	2.2113	0.6521	2.0791
9	0.4077	3.1713	0.5000	2.7301	0.5562	2.4843	0.5925	2.2169	0.6491	2.1297
10	0.4024	3.2275	0.4962	2.7719	0.5518	2.5069	0.5880	2.2393	0.6472	2.1553
20	0.3486	3.5220	0.4395	3.2248	0.5232	2.7649	0.5633	2.4762	0.6228	2.3591
30	0.3057	4.0758	0.4203	0.5425	0.5083	3.0566	0.5348	2.5192	0.5973	2.3924
40	0.2952	4.3930	0.3891	3.6392	0.4958	3.1190	0.5131	2.6812	0.5904	2.5104
50	0.2793	4.4353	0.3874	3.7004	0.4843	3.1396	0.5121	2.7258	0.5870	2.5539

k	n							
	7		8		9		10	
1	0.7698	1.6077	0.7891	1.5732	0.8069	1.5495	0.8201	1.5341
2	0.7335	1.7229	0.7605	1.6883	0.7728	1.6479	0.7835	1.6446
3	0.7123	1.8107	0.7444	1.7368	0.7550	1.7385	0.7666	1.6908
4	0.7018	1.8610	0.7310	1.5922	0.7477	1.7852	0.7558	0.7328
5	0.6192	1.9104	0.7250	1.8373	0.7378	1.8349	0.7467	1.7653
6	0.6841	1.5945	0.7181	1.8821	0.7316	1.8768	0.7398	1.7819
7	0.7104	1.9209	0.7298	1.9051	0.7338	1.8111	0.7552	1.5561
8	0.6744	2.0342	0.6980	1.9459	0.7190	1.9222	0.7306	1.8189
9	0.6711	2.0588	0.6964	1.9618	0.7152	1.9513	0.7281	1.8303
10	0.6693	2.0766	0.6943	1.9997	0.7093	1.9538	0.7224	1.8422
20	0.6504	2.2074	0.6809	2.1062	0.6810	2.0959	0.6985	1.9428
30	0.6345	2.3967	0.6672	2.2398	0.6731	2.1668	0.6848	2.0243
40	0.6109	2.4582	0.6621	2.3577	0.6608	2.1886	0.6739	2.0624
50	0.6092	2.5502	0.6585	2.3803	0.6519	2.2144	0.6729	2.0980

EXTREME VALUE CHARTS AND ANOM BASED ON LOG LOGISTIC DISTRIBUTION

Table 3.8: LLD Constants for Analysis of Means for $b = 5$, $(1-\alpha) = 0.95$

k	n									
	2		3		4		5		6	
1	0.6183	1.7596	0.7003	1.6595	0.7380	1.5590	0.7661	1.5036	0.7871	1.4560
2	0.5683	1.9464	0.6561	1.8041	0.6999	1.6716	0.7287	1.6030	0.7602	1.5422
3	0.5418	2.0700	0.6306	1.8956	0.6762	1.7528	0.7123	1.6647	0.7428	1.5898
4	0.5224	2.1907	0.6141	1.9736	0.6657	1.8008	0.6928	1.7119	0.7313	1.6292
5	0.5175	2.2863	0.6031	2.0374	0.6534	1.8694	0.6816	1.7496	0.7201	1.6586
6	0.5076	2.3282	0.5957	2.0855	0.6480	1.9026	0.6732	1.7694	0.7145	1.6720
7	0.5044	2.3661	0.5802	2.1047	0.6395	1.9335	0.6691	1.7829	0.7093	1.6910
8	0.4913	2.4061	0.5747	2.1234	0.6248	1.9751	0.6587	1.7873	0.7057	1.7134
9	0.4858	2.4285	0.5708	2.1327	0.6195	2.0120	0.6550	1.8075	0.7002	1.7293
10	0.4778	2.4406	0.5634	2.1710	0.6175	2.0183	0.6414	1.8256	0.6942	1.7310
20	0.4292	2.6741	0.5022	2.3900	0.5846	2.1100	0.6218	1.9631	0.6724	1.8581
30	0.3870	2.9332	0.4963	2.6166	0.5771	2.2720	0.5977	2.0171	0.6607	1.9060
40	0.3741	2.9979	0.4683	2.6941	0.5573	2.3353	0.5760	2.0587	0.6469	1.9297
50	0.3563	3.0159	0.4598	2.7000	0.5521	2.3364	0.5729	2.0806	0.6426	1.9329

k	n							
	7		8		9		10	
1	0.8032	1.4224	0.8184	1.3976	0.8342	1.3850	0.8433	1.3707
2	0.7721	1.4942	0.7943	1.4671	0.8050	1.4502	0.8148	1.4294
3	0.7546	1.5472	0.7805	1.5109	0.7900	1.4973	0.8000	1.4701
4	0.7436	1.5846	0.7699	1.5274	0.7838	1.5264	0.7906	1.4873
5	0.7366	1.6084	0.7647	1.5548	0.7760	1.5550	0.7837	1.5172
6	0.7297	1.6367	0.7592	1.5695	0.7694	1.5742	0.7753	1.5216
7	0.7256	1.6484	0.7497	1.5892	0.7669	1.5986	0.7737	1.5360
8	0.7213	1.6674	0.7437	1.6120	0.7580	1.6170	0.7684	1.5413
9	0.7154	1.6722	0.7402	1.6340	0.7507	1.6787	0.7655	1.5481
10	0.7131	1.6751	0.7396	1.6457	0.7500	1.6457	0.7616	1.5503
20	0.6973	1.7975	0.7269	1.7162	0.7264	1.6930	0.7450	1.6237
30	0.6866	1.8671	0.7179	1.8126	0.7159	1.7138	0.7351	1.6402
40	0.6674	1.9378	0.7092	1.8794	0.7115	1.7430	0.7326	1.6644
50	0.6673	1.9386	0.7085	1.8884	0.7012	1.7599	0.7163	1.6892

Example 1 (Wadsworth, 1986)

The following 25 observations are from a metal product manufacturing site. Variations in iron content were suspected in raw material supplied by 5 different suppliers. Five ingots were randomly selected from each of the suppliers. The following table contains the iron determinations for each ingot by weight from each of the 5 suppliers.

Ingot Iron Content (g)	Supplier				
	1	2	3	4	5
	3.46	3.59	3.51	3.38	3.29
	3.48	3.46	3.64	3.4	3.46
	3.56	3.42	3.46	3.37	3.37
	3.39	3.49	3.52	3.46	3.32
	3.4	3.5	3.49	3.39	3.38

The following concentrations were obtained:

Catalyst			
1	2	3	4
58.2	56.3	50.1	52.9
57.2	54.5	54.2	49.9
58.4	57.0	55.4	50.0
55.8	55.3	54.9	51.7

The goodness of fit of data, as revealed by a Q-Q plot (correlation coefficient), for the 3 examples are summarized Table 3.9, which shows that the log logistic distribution is a better model than the normal because it exhibits a significant linear relation between sample and population quantiles.

Example 2

In a study of 3 brands of batteries, it was suspected that the life (in weeks) of the three brands was different. Five of each brand of battery were tested with the following results:

Battery Life (weeks)		
Brand 1	Brand 2	Brand 3
100	76	108
96	80	100
92	75	96
96	84	98
92	82	100

Table 3.9: Goodness of Fit Data from Q-Q Plot

Example	b	LLD	Normal
1	2	0.9306	0.2067
	3	0.9673	
	4	0.9801	
	5	0.9854	
2	2	0.8484	0.4149
	3	0.8986	
	4	0.9206	
	5	0.9324	
3	2	0.8424	0.2067
	3	0.8981	
	4	0.9223	
	5	0.9351	

Example 3

Four catalysts that may affect the concentration of a component in a three component liquid mixture were investigated.

Treating the observations in the data as single samples, the decision limits for the normal and the LLD populations were calculated and are shown in Tables 3.10 and 3.11 respectively.

EXTREME VALUE CHARTS AND ANOM BASED ON LOG LOGISTIC DISTRIBUTION

Table 3.10: Normal Distribution

Examples	[LDL, UDL] (Ott, 1967)	Number of Counts			
		In	P = in/k	Out	Out/k
1: n=5, k=5, $\alpha=0.05$	[3.517, 3.879]	3	0.6	2	0.4
2: n=5, k=3, $\alpha=0.05$	[87.82, 95.52]	2	0.7	1	0.3
3: n=4, k=4, $\alpha=0.05$	[26.14, 82.84]	2	0.5	2	0.5

Table 3.11: Log Logistic Distribution

Examples	[LDL, UDL]	Number of Counts			
		In	P=in/k	Out	Out/k
b = 2					
1 n=5, k=5, $\alpha=0.05$	[1.5345, 25.8322]	5	1	0	0
2 n=5, k=3, $\alpha=0.05$	[44.1805, 572.4044]	3	1	0	0
3 n=4, k=4, $\alpha=0.05$	[22.2330, 420.2953]	4	1	0	0
b = 2					
1 n=5, k=5, $\alpha=0.05$	[1.9053, 10.4989]	5	1	0	0
2 n=5, k=3, $\alpha=0.05$	[54.3510, 259.3351]	3	1	0	0
3 n=4, k=4, $\alpha=0.05$	[28.3547, 180.0506]	4	1	0	0
b = 3					
1 n=5, k=5, $\alpha=0.05$	[2.1685, 7.2966]	5	1	0	0
2 n=5, k=3, $\alpha=0.05$	[60.8822, 183.1867]	3	1	0	0
3 n=4, k=4, $\alpha=0.05$	[32.8685, 120.9012]	4	1	0	0
b = 5					
1 n=5, k=5, $\alpha=0.05$	[2.3499, 6.0319]	5	1	0	0
2 n=5, k=3, $\alpha=0.05$	[65.2987, 152.5993]	3	1	0	0
3 n=4, k=4, $\alpha=0.05$	[36.2766, 98.1243]	4	1	0	0

Conclusion

Ott's (1967) ANOM tables are designed for normal distributions, the number of homogenous means for each data set was 3, 2, 2, respectively, and those not homogeneous are 2, 1 and 2, respectively. When the ANOM tables of the proposed model (LLD) are used for the same data sets, the number of homogenous means are 5, 3 and 4, respectively, and they do not exhibit deviation from homogeneity for values of $b = 2$, $b = 3$, $b = 4$ and $b = 5$. Use of the normal model resulted in homogeneity for some means and deviation for other means, thus indicating possible rejection of those means. The rejection decision is valid if a normal distribution is a good fit for the data. However, by comparison, results show that the LLD is a better model than the normal. Results are supported by the Q-Q plot correlation coefficient for each data set with the normal and with the LLD. It is therefore assumed that more error is likely to be associated with the decision process when data are from a normal distribution, thus, making all the means homogenous using LLD (see Table 3.11) is recommended over using the normal-ANOM procedure.

References

- Bakir, S. T. (1994). Means using the ranks for randomized complete block designs. *Communications in Statistics-Simulation and Computation*, 23, 547-568.
- Bernard, A. J., & Wludyka, P. S. (2001). Robust I-sample analysis of means type randomization tests for variance. *Journal of Statistical Computation and Simulation*, 69, 57-88.
- Farnum, N. R. (2004). Analysis of means table using mathematical processors. *Quality Engineering*, 16, 399-405.
- Guirguis, G. H., & Tobias, R. D. (2004). On the computation of the distribution for the analysis of means. *Communication in Statistics-Simulation and Computation*, 16, 861-887.
- Montgomery, D. C. (2000). *Design and Analysis of Experiments*, 5th Ed. New York, NY: John Wiley and Sons Inc.
- Nelson, P. R., & Dudewicz, E. J. (2002). Exact analysis of means with unequal variances. *Technometrics*, 44, 152-160.
- Ott, E. R. (1967). Analysis of means: A graphical procedure. *Industrial Quality Control*, 24, 101-109.
- Ramig, P. F. (1983). Applications of analysis of means. *Journal of Quality Technology*, 15, 19-25.
- Rao, C. V. (2005). Analysis of means: A review. *Journal of Quality Technology*, 37, 308-315.
- Rao, C. V., & Prankumar, M. (2002). ANOM-type graphical methods for testing the equality of several correlation coefficients. *Gujarat Statistical Review*, 29, 47-56.
- Shewart, W. A. (1986). *Statistical method from the view point of quality control*, 14th Ed. New York, NY: Mineola, Dover Publications.
- Srinivasa Rao, B., & Kantam, R. R. L. (2012). Extreme value charts and analysis of means based on half logistic distribution. *International Journal of Quality, Reliability and Management*, 29(5), 501-511.
- Wadsworth, H. M., Stephens, K. S., & Godfrey, A. F. (2001). *Modern methods of quality control and improvement*. New York, NY: John Wiley and Sons, Inc.
- Wludyka, P. S., Nelson, P. R., & Silva, P. R. (2001). Power curves for analysis of means for variances. *Journal of Quality Technology*, 3, 60-75.

Ferrieri's Index of Openness Applied to Remittances to Developing Countries

Gaetano Ferrieri
Studi Interdisciplinari,
AISI, BA, Italy

A new methodology to measure international openness and globalization is described. This allows capacity to be effectively combined with size in a number of socio-economic areas, such as trade, migration and foreign investment. The method is applied to remittances to developing countries.

Key words: Openness index, statistical methods, remittances, globalization.

Introduction

A new method to measure international openness, hereafter termed the Ferrieri's Index of Openness (FIO), consists of a synthetic indicator to measure the capacity of countries for various socio-economic phenomena adjusted by weight and including the influence of other key related aggregates, such as population or gross domestic product (Ferrieri, 2010; 2006). The FIO has been applied to a number of transferable socio-economic phenomena, for example, trade, foreign direct investment and migration. This article applies the method to analyzing remittances to developing countries and demonstrates it using sample calculations and detailed technical observations.

Overview

Compared with previous work with the FIO (Ferrieri 2010), the innovative methodology is applied to another macroeconomic context by analyzing remittances to developing countries. It is further shown its effectiveness in providing a more comprehensive approach for measuring distinctly relative and absolute dimensions.

Data regarding inflow remittances are employed to developing countries and sample calculations and detailed technical observations are suitably provided. Some advice regarding the elaboration of basic data is also provided and applied, for example, use of three-year averages in order to reduce yearly fluctuations in the main aggregate. Data collection, preparation and normalization for capacity effect and adjusting for size effect are illustrated. In addition, the choice of coefficient of variation (CV or k factor) of the denominator of the basic indicator is explained along with the index range (0-1), including the specific role of the coefficient of variation (k). Illustrations (Tables 3a and 3b) show the top performers in terms of difference in value between synthetic index and normalized indicator, and in terms of rank. The benefits and limits when reducing the maximum reachable size from 100% of the total of the countries (standard or basic scenario) to a lower proportion are described. Finally, further technical observations on the index formulation and applications are provided.

Methodology

Remittances are defined as the sum of workers' remittances, compensation of employees and migrant transfers (World Bank, 2010). Together with foreign direct investment and official development aid, they represent a key financial source for developing countries. International remittance data are typically expressed in US dollars and are managed and published by the World Bank (World Bank, 2011). Like other macroeconomic indicators, statistical measures

Gaetano Ferrieri (PhD) is the founding director of AISI, an independent think-tank of interdisciplinary studies devoted to International Development (Italy). His research interests include: measuring methods (e.g., synthetic indicators), global integration, international relations and labor markets. Email him at: ferrieri@aisigf.org (1); gaetanoferrieri@tiscali.it (2).

related to this aggregate are provided in both absolute and relative terms, often as a percentage of the gross domestic product (GDP) in receiving countries. Both absolute and relative perspectives provide useful, but distinct, snapshots of the phenomenon. Size represents a key factor in influencing the results provided by the sole relative approach – that is, remittances as proportion of GDP in receiving countries – impeding big economies (e.g., China, India) to match the same performance in terms of capacity of smaller ones (e.g., Tajikistan, Lesotho). Thus, there is a need to compare countries by following a more comprehensive approach that adequately considers and combines capacity and size in order to reduce the gap between big and small economies, to recognize the importance of size (and related factors) and to preserve the role of capacity. Based on its formulation and ability to consider a wide range of applications to transferable phenomena, Ferrieri's Index of Openness (FIO) appears to offer a suitable and effective methodological tool for this purpose.

The FIO is a mathematical function that combines the capacity of countries for a given transferable phenomenon with their share in the same, taking into consideration the influence of other key related aggregates, for example, gross domestic product (GDP). Similar to other transferable phenomena, the FIO can be applied to both inflow and outflow remittances. This article focuses on remittance inflows because this issue seems to be more consistent with the macroeconomic profile and situations of developing countries. Analogous to other phenomena, such as trade and migration, the FIO calculation methodology applied to remittances is articulated in two phases.

Phase 1

Data related to aggregates to be analyzed must be collected; in this case, inflow remittances (REM) to developing countries and their gross domestic product (GDP). Remittance data used in this study are from the World Bank and GDP data are from the International Monetary Fund (IMF). Data are expressed in US dollars at market exchange rates. Only countries with available and comparable data (both REM and GDP) in the given time horizon are

considered. These data were used to elaborate the basic indicator REM-to-GDP ratio. Because remittances, like other macroeconomic aggregates, can fluctuate from year to year, three-year averages were calculated for preparing the basic indicator REM-to-GDP ratio. The REM-to-GDP ratio was then elaborated for all countries to be monitored and analyzed. Although the first two decimals can be retained for illustration purposes (tables, graphs, etc.), all figures are considered in electronic calculations in order to better define their precise ranking.

The indicator values were normalized on a scale to one, in which unity corresponds to the highest value across all countries analyzed. In this work, the benchmark is the maximum value at the current data point (three-year average: 2008-10). In order to determine time comparisons without needing to index recalculations, it is suitable to fix the highest value observed over time or a given time horizon as the benchmark (Ferrieri 2010; 2006).

Phase 2

Phase 2 consists in adjusting the country indicator values normalized for weight of the country in the total aggregate, which is their total remittances, while at the same time taking into consideration the dispersion of the denominator of the basic indicator, the GDP. This second step starts by calculating the weight or share (not in percentage terms) of each country in the total aggregate (remittances). Note that, although only up to three decimals are shown in illustrations, all decimals are (and should be) considered in electronic calculations. These weights are then subtracted from one, when unity corresponds to the theoretical maximum share (total of countries in the standard approach). These calculated differences are then raised to the coefficient of variation ($CV = \text{standard deviation divided by mean}$) of the aggregate defined by the denominator of the basic indicator: in this case GDP. This factor measures the relative dispersion of the second aggregate expressing the basic indicator.

As noted, the denominator is very important in determining the basic indicator value. Until 2006, this second aggregate was considered the first exponent in the FIO formula because its statistical influence was considered

to be similar to that of the main aggregate (numerator). Accordingly, the weight of the countries in total GDP was calculated and this share subtracted from one; this was then multiplied by the difference from one related to the key aggregate (numerator) and the result was considered the exponent of the power function in which the base was the normalized indicator value (Ferrieri, 2006).

To better identify the role played by the main aggregate (in this case, remittances), the exponent of the FIO formula was redefined by expressing the statistical importance of the denominator (the second aggregate), GDP for example, in terms of dispersion. This factor has the following properties:

1. It continues to express the importance of the aggregate at the denominator in terms of dispersion (relative variability);
2. It is constant for all countries in order to better appreciate the changes in the main aggregate; and
3. It contributes to coherent determination of the impact of size.

The coefficient of variation (CV) is the best empirical indicator to comply with all properties and needs; a higher CV indicates a higher (relative) variability of a given related phenomenon at the denominator (in this case, GDP). Being a constant factor for all countries to be compared, a higher CV mainly benefits those countries with a greater size in the phenomenon concerned; in other words, it amplifies the size effect for all countries, but particularly for those having a higher weight in the phenomenon analyzed.

The formula for Ferrieri's Index of Openness (FIO) is:

$$\text{Index} = \left(\frac{V_i}{V_{\text{MAX}}} \right)^{(1-\Pi)^{\kappa}} \quad (1)$$

where, considering the specific phenomenon analyzed (remittances), V_i is the value of the basic indicator (in this study: remittances-to-

GDP ratio) for each country in the given time; V_{MAX} is the maximum value of the basic indicator across the countries; Π is the share of each country in the world aggregate considered (in this study: remittance inflows) in the given time, not expressed in percentage terms; κ is the coefficient of variation of the denominator (in this study: GDP) calculated over the countries analyzed in the given time, not expressed in percentage terms.

The two different effects determining the FIO value are defined respectively as capacity effect and size effect. These are calculated as:

$$\text{Capacity Effect} = V_i / V_{\text{MAX}}$$

and

$$\text{Size Effect} = (1 - \Pi)^{\kappa}$$

where exponent κ = constant.

The index value is determined by the capacity effect (base of the power), when the size effect (exponent of the power) implies a growth in the index value for all countries, as much higher as their share in the phenomenon concerned (Ferrieri, 2010; 2006). The maximum index value is one and a country can realize this score in two ways:

1. By matching the best capacity (highest indicator value), or mathematically: $V_i = V_{\text{MAX}}$, therefore $V_i / V_{\text{MAX}} = 1$ and $\text{FIO} = 1$; or
2. By monopolizing the whole phenomenon or reaching the theoretical best size. It should be noted that, although this latter hypothesis is both unlikely and unrealistic, it should be retained in mathematical terms. Under this (extreme) hypothesis, mathematically: $(1 - \Pi)^{\kappa} = (1-1)^{\kappa} = 0$, therefore $\text{FIO} = 1$.

As observed, the size effect is also determined by the factor κ . Because this exponent is equal for all countries, the most benefited countries are those with a larger size (Π). As intuitively understandable, a higher κ increases the size effect, particularly for larger size countries. The best performer in terms of capacity is not influenced by any change in the κ

factor for the same reasons why the power function (as mathematically formulated) cannot improve or worsen a situation given by the best capacity (unity).

Two extreme cases are possible in this regard. If κ is equal to one (GDP standard deviation = GDP mean), the index value for all countries is determined by their capacity effect and a size effect based only on their share in the phenomenon concerned. If (paradoxically) κ is equal to zero (GDP standard deviation = 0), meaning GDP is the same for all countries (there is no variability), then the index value is only given by the capacity effect and this seems to be consistent with openness (Ferrieri, 2010). In such extreme cases, the difference between countries is given only by their basic components.

Results

The FIO was calculated over 118 developing countries with available data in both relevant aggregates: remittances and GDP. Three-year averages were calculated in order to adjust for yearly fluctuations; however, 2011 data were not considered because they were still estimations. Remittances were reported to GDP in order to build the basic indicator resulting in a remittances-to-GDP ratio for the three year average (2008-10). The countries' values for this indicator were reported to the highest value across the same countries compared (in this study: Tajikistan: 41.56%) in order to have normalized values referring to one; this normalized indicator represents the capacity for the given phenomenon in a comparative approach. This indicator of capacity (base) was then raised to the size effect, which was calculated as the distance from one of each country's share in total remittances raised to the GDP coefficient of variation. Table 1 provides sample calculations referring to China, India and Tajikistan. Results for all countries are shown in Table 2. The FIO index values applied to remittances is conventionally defined as IOREM.

As shown in the tables, the highest indicator (REM-to-GDP ratio) value across countries compared is that of Tajikistan (41.56%), thus the indicator value normalized to Tajikistan corresponds to the benchmark (unity).

India and China are respectively the first and second by share in total remittances, by representing respectively 16.1% and 15.8% of the total remittances among the 118 countries analyzed; without considering the size effect they would rank 58th and 86th out of the 118 developing countries. By taking into account the size effect, their IOREM values rise to 0.277 for India and 0.133 for China. The growth, in terms of index value, for India is 210.91% and for China is 464.19%; the size effect allows India to gain 38 positions in ranking (from 58th to 20th) and China 43 places (from 86th to 43rd). Understandably, value and rank remain unchanged for Tajikistan, which is the best performer.

It is important to emphasize that the size effect causes index values to increase for all countries – most notably for those with higher size in the related phenomenon (i.e., remittances). The last two columns in Table 2 show rank by indicator value normalized (IVN) and IOREM (index value combining capacity with size), out of the 118 countries analyzed: 20 of them (about one sixth) improve in rank, 63 (more than half) decline in rank and 35 (less than one third) remain unchanged in their position.

The key factor determining the performances of countries is their capacity, particularly when their size is similar or not significantly different. For example, Haiti and Lithuania have a similar size in total remittances inflow, but the indicator value of Haiti is six times higher than that of Lithuania. Due to the size effect (Table 2, third column), both countries (like all others) gain in terms of value, but the higher capacity of Haiti compared to Lithuania allows Haiti to lose just one position passing from IVN to IOREM, while Lithuania loses three places. Conversely, size fosters changes in ranking between countries when their capacity is somewhat similar. For example, Lebanon has a slightly lower indicator value compared to Haiti (21.39% versus 21.57%), but due to the size effect, Lebanon gains one position compared to Haiti in the IOREM ranking. A similar situation is observed for Albania and Bangladesh: the indicator value (and so the indicator value normalized) of Albania is slightly higher than that of

FERRIERI'S INDEX OF OPENNESS FOR DEVELOPING COUNTRY REMITTANCES

Bangladesh: 10.71% compared to 10.64% (or 0.258 compared to 0.256 in terms of normalized indicators). Due to the size effect, Albania loses two positions and Bangladesh gains three places in ranking (in terms of value, the IOREM of Albania is 0.263 and that of Bangladesh is 0.297).

Table 3a shows the top 20 gainers in terms of difference in value between IOREM and IVN. Apart from China, India and Mexico, which improve exceptionally in both value and rank, other countries show high performances in terms of value but not necessarily in terms of rank. For example, Russia's index improves by 33.47% but its rank improves by just one position. By contrast, Egypt improves by 22.02% in terms of index value (less than Russia) but gains five places in rank. This is mainly due (taking also into account the

different sizes of the countries) to the different capacity: the indicator value of Russia is much lower than that of Egypt (see Table 2). In another example, Ukraine with a growth of 15.70% – less than half compared to that of Russia – also gains one position. The size effect for these two countries is not dissimilar; the real difference is due to their very different capacities.

Table 3b shows the top 20 gainers in terms of rank. The first three places between the two classifications (Table 3a and 3b) are the same, however, for newcomers like Lebanon, Azerbaijan, Jordan, Kazakhstan and the Kyrgyz Republic a slower increase in terms of value is sufficient to cause a gain in ranking comparable to that of other better performing countries in terms of value.

Table 1: Ferrieri's Index of Openness Applied to Remittances (IOREM) to Developing Countries*
Sample Calculations: China, India and Tajikistan (2008-10)

Variables	China	India	Tajikistan
Remittances-to-GDP ratio (V_i)	0.98 %	3.70 %	41.56 %
(A) Capacity Effect: $(IVN = V_i/V_{MAX})^{**}$	0.024	0.089	1.000
IVN (or V_i) rank	86	58	1
Share in total remittances (Π)***	0.158	0.161	0.007
Constant = coefficient of variation of GDP (κ)****	3.59	3.59	3.59
(B) Size Effect: $(1 - \Pi)^\kappa$	0.539	0.531	0.975
Index of Openness to Remittances (IOREM) = $(A)^{(B)}$	0.133	0.277	1.000
IOREM rank	43	20	1
Difference between IOREM and IVN value	464.19 %	210.91 %	-
Difference between IOREM and IVN rank	43	38	-

*REM: Remittances (US dollars at market exchange rates). GDP: (Nominal) Gross Domestic Product (US dollars at market exchange rates). Values refer to three-year averages 2008-10. Although index values are expressed up to three decimal points their ranks reflect all significant figures. Source: World Bank (2011) and IMF (2011). ** V_{MAX} is the maximum value of V_i across the 118 countries analysed in the given time and corresponding to 41.56% (Tajikistan); ***The share is calculated on the total of 118 developing countries with available data; ****The coefficient of variation of GDP (κ) is calculated over the 118 world economies analysed.

Table 2: Index of Openness to Remittances (IOREM): Developing Countries (2008-10)*

Country	REM/GDP% Vi	Share in total REM	Size Effect	Value		Rank	
				IVN	IOREM	IVN	IOREM
Albania	10.71	0.004	0.985	0.258	0.263	21	23
Algeria	1.35	0.007	0.976	0.032	0.035	79	80
Antigua and Barbuda	1.93	0.000	1.000	0.046	0.046	71	74
Argentina	0.19	0.002	0.993	0.005	0.005	106	106
Armenia	9.53	0.003	0.989	0.229	0.233	24	25
Azerbaijan	2.96	0.004	0.984	0.071	0.074	65	64
Bangladesh	10.64	0.032	0.890	0.256	0.297	22	19
Belarus	0.71	0.001	0.996	0.017	0.018	92	92
Belize	5.79	0.000	0.999	0.139	0.140	38	41
Benin	3.78	0.001	0.997	0.091	0.092	55	61
Bhutan	0.33	0.000	1.000	0.008	0.008	100	100
Bolivia	6.13	0.003	0.988	0.147	0.151	36	38
Bosnia & Herzegovina	13.00	0.007	0.975	0.313	0.322	16	17
Botswana	0.75	0.000	0.999	0.018	0.018	91	91
Brazil	0.25	0.014	0.951	0.006	0.008	105	103
Bulgaria	3.31	0.005	0.982	0.080	0.083	61	63
Burkina Faso	1.15	0.000	0.999	0.028	0.028	83	84
Burundi	1.51	0.000	1.000	0.036	0.036	77	79
Cambodia	3.06	0.001	0.996	0.074	0.074	63	65
Cameroon	0.81	0.001	0.998	0.020	0.020	90	90
Cape Verde	8.93	0.000	0.998	0.215	0.215	25	26
Chile	0.00	0.000	1.000	0.000	0.000	118	118
China	0.98	0.158	0.539	0.024	0.133	86	43
Colombia	1.73	0.014	0.951	0.042	0.049	74	72
Congo, Rep.	0.13	0.000	1.000	0.003	0.003	112	112
Costa Rica	1.76	0.002	0.994	0.042	0.043	73	75
Côte d'Ivoire	0.82	0.001	0.998	0.020	0.020	88	88
Djibouti	3.02	0.000	1.000	0.073	0.073	64	67
Dominica	5.48	0.000	1.000	0.132	0.132	40	44
Dominican Republic	7.30	0.011	0.961	0.176	0.188	30	32
Ecuador	4.81	0.008	0.970	0.116	0.123	46	46
Egypt, Arab Rep.	4.14	0.025	0.914	0.100	0.122	52	47
El Salvador	16.76	0.011	0.960	0.403	0.418	12	12
Ethiopia	0.99	0.001	0.997	0.024	0.024	85	86
Fiji	4.82	0.000	0.998	0.116	0.117	45	49

*Notes: REM: Remittances (US dollars at market exchange rates). GDP: (Nominal) Gross Domestic Product (US dollars at market exchange rates). Values refer to three-year averages 2008-10. IVN: Indicator value normalized: $V_i/V_{MAX} = 41.56$. Size Effect: calculated by raising the difference from one values in the second column (share in total) to the k value = 3.594 (coefficient of variation of GDP). Although index values are expressed up to three decimal points, ranks reflect all significant figures. Source: World Bank (2011) and IMF (2011).

FERRIERI'S INDEX OF OPENNESS FOR DEVELOPING COUNTRY REMITTANCES

Table 2 (continued): Index of Openness to Remittances (IOREM): Developing Countries (2008-10)*

Country	REM/GDP% Vi	Share in total REM	Size Effect	Value		Rank	
				IVN	IOREM	IVN	IOREM
Gambia, The	8.46	0.000	0.999	0.204	0.204	26	27
Georgia	6.38	0.002	0.992	0.154	0.156	35	35
Ghana	0.43	0.000	0.999	0.010	0.011	96	97
Grenada	6.86	0.000	0.999	0.165	0.165	33	34
Guatemala	10.77	0.013	0.953	0.259	0.276	19	21
Guinea	1.43	0.000	0.999	0.034	0.034	78	81
Guinea-Bissau	5.84	0.000	0.999	0.141	0.141	37	40
Guyana	13.96	0.001	0.997	0.336	0.337	15	16
Haiti	21.57	0.004	0.984	0.519	0.524	8	9
Honduras	18.54	0.008	0.970	0.446	0.457	10	10
India	3.70	0.161	0.531	0.089	0.277	58	20
Indonesia	1.17	0.022	0.925	0.028	0.037	82	78
Iran, Islamic Rep.	0.30	0.004	0.987	0.007	0.008	104	104
Iraq	0.09	0.000	0.999	0.002	0.002	115	115
Jamaica	15.42	0.006	0.977	0.371	0.380	13	14
Jordan	15.26	0.012	0.959	0.367	0.383	14	13
Kazakhstan	0.19	0.001	0.997	0.004	0.005	108	107
Kenya	5.56	0.005	0.981	0.134	0.139	39	42
Kosovo	17.69	0.003	0.989	0.426	0.430	11	11
Kyrgyz Republic	24.25	0.004	0.987	0.584	0.588	6	5
Lao PDR	0.55	0.000	1.000	0.013	0.013	95	95
Lebanon	21.39	0.023	0.918	0.515	0.543	9	8
Lesotho	34.28	0.002	0.993	0.825	0.826	2	2
Liberia	4.04	0.000	1.000	0.097	0.097	54	59
Libya	0.02	0.000	1.000	0.001	0.001	117	117
Lithuania	3.51	0.004	0.984	0.085	0.088	59	62
Macedonia, FYR	4.13	0.001	0.996	0.099	0.101	53	57
Malaysia	0.58	0.004	0.986	0.014	0.015	94	94
Maldives	0.19	0.000	1.000	0.004	0.004	107	108
Mali	4.86	0.001	0.995	0.117	0.118	44	48
Mauritius	2.31	0.001	0.998	0.056	0.056	69	70
Mexico	2.33	0.074	0.759	0.056	0.112	68	52
Moldova	25.88	0.005	0.983	0.623	0.628	4	4
Mongolia	4.27	0.001	0.997	0.103	0.103	50	56
Montenegro	7.08	0.001	0.997	0.170	0.172	32	33

*Notes: REM: Remittances (US dollars at market exchange rates). GDP: (Nominal) Gross Domestic Product (US dollars at market exchange rates). Values refer to three-year averages 2008-10. IVN: Indicator value normalized: $V_i/V_{MAX} = 41.56$. Size Effect: calculated by raising the difference from one values in the second column (share in total) to the k value = 3.594 (coefficient of variation of GDP). Although index values are expressed up to three decimal points, ranks reflect all significant figures. Source: World Bank (2011) and IMF (2011).

Table 2 (continued): Index of Openness to Remittances (IOREM): Developing Countries (2008-10)*

Country	REM/GDP% Vi	Share in total REM	Size Effect	Value		Rank	
				IVN	IOREM	IVN	IOREM
Morocco	7.23	0.021	0.928	0.174	0.197	31	31
Mozambique	1.22	0.000	0.999	0.029	0.029	80	82
Myanmar	0.33	0.000	0.999	0.008	0.008	101	101
Namibia	0.14	0.000	1.000	0.003	0.003	111	111
Nepal	22.33	0.010	0.966	0.537	0.549	7	7
Nicaragua	12.59	0.003	0.991	0.303	0.306	17	18
Niger	1.71	0.000	0.999	0.041	0.041	75	76
Nigeria	5.12	0.031	0.892	0.123	0.154	41	37
Pakistan	5.06	0.027	0.907	0.122	0.148	43	39
Panama	0.82	0.001	0.998	0.020	0.020	89	89
Papua New Guinea	0.16	0.000	1.000	0.004	0.004	109	110
Paraguay	3.77	0.002	0.993	0.091	0.092	57	60
Peru	1.81	0.008	0.972	0.044	0.048	72	73
Philippines	11.05	0.063	0.792	0.266	0.350	18	15
Romania	3.44	0.019	0.933	0.083	0.098	60	58
Russian Federation	0.38	0.018	0.938	0.009	0.012	97	96
Rwanda	1.63	0.000	0.999	0.039	0.039	76	77
Samoa	26.30	0.000	0.998	0.633	0.633	3	3
São Tomé & Príncipe	1.14	0.000	1.000	0.027	0.027	84	85
Senegal	10.73	0.004	0.984	0.258	0.264	20	22
Serbia	7.94	0.011	0.963	0.191	0.203	27	28
Seychelles	1.17	0.000	1.000	0.028	0.028	81	83
Sierra Leone	2.31	0.000	1.000	0.056	0.056	70	71
Solomon Islands	0.35	0.000	1.000	0.008	0.008	99	99
South Africa	0.31	0.003	0.989	0.007	0.008	102	102
Sri Lanka	7.91	0.011	0.961	0.190	0.203	28	29
St. Kitts & Nevis	6.43	0.000	1.000	0.155	0.155	34	36
St. Lucia	2.72	0.000	1.000	0.065	0.066	67	69
St. Vincent & Grenadines	4.47	0.000	1.000	0.108	0.108	48	53
Sudan	4.15	0.008	0.973	0.100	0.106	51	54
Suriname	0.11	0.000	1.000	0.003	0.003	113	113
Swaziland	3.07	0.000	0.999	0.074	0.074	62	66
Syrian Arab Republic	2.77	0.005	0.983	0.067	0.070	66	68
Tajikistan	41.56	0.007	0.975	1.000	1.000	1	1
Tanzania	0.11	0.000	1.000	0.003	0.003	114	114

*Notes: REM: Remittances (US dollars at market exchange rates). GDP: (Nominal) Gross Domestic Product (US dollars at market exchange rates). Values refer to three-year averages 2008-10. IVN: Indicator value normalized: $V_i/V_{MAX} = 41.56$. Size Effect: calculated by raising the difference from one values in the second column (share in total) to the k value = 3.594 (coefficient of variation of GDP). Although index values are expressed up to three decimal points, ranks reflect all significant figures. Source: World Bank (2011) and IMF (2011).

FERRIERI'S INDEX OF OPENNESS FOR DEVELOPING COUNTRY REMITTANCES

Table 2 (continued): Index of Openness to Remittances (IOREM): Developing Countries (2008-10)*

Country	REM/GDP% Vi	Share in total REM	Size Effect	Value		Rank	
				IVN	IOREM	IVN	IOREM
Thailand	0.62	0.006	0.980	0.015	0.016	93	93
Togo	10.55	0.001	0.996	0.254	0.255	23	24
Tonga	24.28	0.000	0.999	0.584	0.585	5	6
Tunisia	4.46	0.006	0.978	0.107	0.113	49	51
Turkey	0.16	0.003	0.988	0.004	0.004	110	109
Uganda	5.11	0.003	0.991	0.123	0.125	42	45
Ukraine	3.78	0.017	0.939	0.091	0.105	56	55
Uruguay	0.30	0.000	0.999	0.007	0.007	103	105
Vanuatu	0.93	0.000	1.000	0.022	0.022	87	87
Venezuela, RB	0.04	0.000	0.998	0.001	0.001	116	116
Vietnam	7.35	0.022	0.923	0.177	0.202	29	30
Yemen, Rep.	4.57	0.004	0.986	0.110	0.114	47	50
Zambia	0.35	0.000	0.999	0.008	0.008	98	98

*Notes: REM: Remittances (US dollars at market exchange rates). GDP: (Nominal) Gross Domestic Product (US dollars at market exchange rates). Values refer to three-year averages 2008-10. IVN: Indicator value normalized: $V_i/V_{MAX} = 41.56$. Size Effect: calculated by raising the difference from one values in the second column (share in total) to the k value = 3.594 (coefficient of variation of GDP). Although index values are expressed up to three decimal points, ranks reflect all significant figures. Source: World Bank (2011) and IMF (2011).

For example, the Kyrgyz Republic needs to grow by only 0.71% in value to gain one rank while Russia must increase its value by 33.47% (or 47 times more than its neighbour) to obtain the same result. The difference can be explained by the much higher capacity of Kyrgyz Republic compared to that of Russia (24.25% compared to 0.38%: about 64 times as much).

Conclusion

Ferrieri's Index of Openness has a wide range of applications in socio-economic fields and – based on its conceptual and mathematical properties – it appears to be a valid statistical tool to analyze remittances. As shown herein, the methodology combines the capacity of countries for a given transferable phenomenon (remittances) with their size in a suitable way by considering the role of size (including any related factor) while at the same time preserving capacity. The index can be calculated on a yearly basis as well as along other time horizons, such as three-yearly basis.

To compare countries over time, a suitable benchmark in terms of capacity must be fixed: this could be the highest indicator value across countries over the given period of time. From this time comparison perspective, it can be assumed that k can be calculated for the current yearly or three-yearly value for the countries to be analyzed. If one wants to appreciate changes in capacity and size, regardless of changes in terms of dispersion in GDP, it is also possible to calculate a k factor over the given period of time (preferably on the basis of appropriate methods like real and/or parity power purchasing terms), by taking into account that the limit of such an approach is to have current GDP data points calculated on nominal yearly (or three-yearly) basis, and GDP variability factor (k or CV) fixed on a longer time horizon. In such a situation $k = 0$ does not necessarily mean equal GDP for the current data points, because GDP is calculated yearly or three-yearly when k is over a longer time.

Table 3a: Index of Openness to Remittances (IOREM)
Differences between IOREM and IVN Values and Ranks (2008-10), Top 20 Gainers in Terms of Value*

Country	IOREM compared to IVN	
	Difference in Value (%)	Difference in Rank
China	464.19	43
India	210.91	38
Mexico	100.04	16
Russian Federation	33.47	1
Philippines	31.80	3
Indonesia	30.90	4
Brazil	28.80	2
Nigeria	25.25	4
Egypt, Arab Rep.	22.02	5
Pakistan	21.61	4
Romania	18.22	2
Colombia	16.76	2
Bangladesh	16.16	3
Ukraine	15.70	1
Vietnam	14.37	-1
Morocco	13.44	-
Peru	9.05	-1
Thailand	8.73	-
Algeria	8.44	-1
Turkey	7.20	1

*IVN: Indicator value normalized: $V_i/V_{MAX} = 41.56$, Source: World Bank (2011) and IMF (2011)

Ferrieri (2010) illustrated the flexibility of his method in allowing a reduction in the maximum reachable size from 100% of the total of the countries (standard or basic scenario) to a lower proportion, such as 25% of the same aggregate. A reduction in the maximum reachable size allows a better balance between capacity and size.

Further details and observations are needed in this regard. Ferrieri (2010) also showed that, in a scenario characterized by a lower reachable size, all countries improve their index value, particularly those having a higher size compared to the standard situation in which the upper limit is the total of the same countries compared. Mathematically this is because $(1 - \Pi)^\kappa$ – when the exponent κ is constant – decreases when the share Π increases (due to the reduction in the reference aggregate or

maximum reachable share). In addition, because the base (V_i/V_{MAX}) – expressed on a scale to one – is raised to a minor distance from one, the final score (index value) is higher. The total of the countries' weights $\Sigma \Pi$ will no longer be 1 (as in the basic approach, where $\Sigma \Pi$ corresponds to the total of countries compared) but will be greater depending on the reducing factor used.

For example, by reducing the maximum reachable size to one-third, $\Sigma \Pi = 3$, to one-fourth is 4, to one-fifth is 5, etc. Also, note that the size effect is determined by the factor k that amplifies the effect.

In summary, a reduction in maximum reachable size allows all countries, and notably those with a high weight, to reduce considerably their distance from the best(s) performer(s). Such a reduction makes the size effect more powerful. Clearly, any change in maximum

FERRIERI'S INDEX OF OPENNESS FOR DEVELOPING COUNTRY REMITTANCES

Table 3b: Index of Openness to Remittances (IOREM)
Differences between IOREM and IVN Values and Ranks (2008-10), Top 20 Gainers in Terms of Rank*

Country	IOREM compared to IVN	
	Difference in Rank	Difference in Value (%)
China	43	464.19
India	38	210.91
Mexico	16	100.04
Egypt, Arab Rep.	5	22.02
Indonesia	4	30.90
Nigeria	4	25.25
Pakistan	4	21.61
Philippines	3	31.80
Bangladesh	3	16.16
Brazil	2	28.80
Romania	2	18.22
Colombia	2	16.76
Russian Federation	1	33.47
Ukraine	1	15.70
Turkey	1	7.20
Lebanon	1	5.58
Azerbaijan	1	4.32
Jordan	1	4.20
Kazakhstan	1	1.53
Kyrgyz Republic	1	0.71

*IVN: Indicator value normalized: $V_i / V_{MAX} = 41.56$, Source: World Bank (2011) and IMF (2011)

reachable size does not compromise the situation of the best performing country, which will steadily remain in the position of highest capacity. This is consistent with the conceptual and mathematical properties of the FIO as described and illustrated.

The choice of scenarios based on a reduction in maximum reachable size depends on the extent to which importance is placed on the size effect. In principle, such a decision is at the discretion of the analyst/researcher making use of the described methodology.

References

Ferrieri, G. (2006). Method for measuring international openness. *Social Indicators Research* 77(2), 245-255.

Ferrieri, G. (2010). Measuring openness. *Journal of Modern Applied Statistical Methods*, 9(1), 172-180.

Ferrieri, G. *Index of international openness*. Reports obtainable from: <http://www.aisigf.org>.

IMF. (2011). *The world economic outlook database*. Washington, DC. <http://www.imf.org>.

World Bank. (2010). *Migration and remittances factbook 2011, 2nd Edition*. Washington DC. <http://www.worldbank.org>.

World Bank. (2011). *Remittances data inflows*. Washington, DC. <http://www.worldbank.org>.

Weighted Cook-Johnson Copula and their Characterizations: Application to Probable Modeling of the Hot Spring Eruptions

Hakim Bekrizadeh Gholam Ali Parham Mohamd Reza Zadkarmi
Shahid Chamran University,
Ahvaz, Iran

Copulas have emerged as a practical method for multivariate modeling. A limited amount of work has been conducted regarding the application of copula-based modeling in context analysis. This study generalizes the Cook-Johnson copula under the appropriate weighted function and provides examples and the properties of the generalized Cook-Johnson copula. Results show that the generalized Cook-Johnson copula is suitable for probable modeling of hot spring eruption.

Key words: Cook-Johnson copula, weighted distribution functions, measures of dependence, hot spring eruption.

Introduction

Research has shown that it is important to consider dependence among variables in studies as opposed to ignoring the dependence structure solely for mathematical simplicity. As De Michele, et al. (2005) showed, among other consequences, failure to take dependence between variables into account may lead to either over- and under- estimation of the parameters of a model. Copulas have emerged as a practical and efficient method for multivariate event analysis (Joe, 1997; Nelsen, 2006). Application of copulas has increased in the hydrological field as evidenced by Genest and Favre (2007), Gebremichael and Krajewski (2007) and others in a special issue of the *Journal of Hydrological Engineering*. An advantage of using copulas is that marginal behaviors and dependence structure can be studied separately. Most copula applications are in bivariate analysis, for example, De Michele and Salvadori (2003; 2004 a, b) applied copula

modeling to examine the dependence between storm intensity and duration. In addition, Zhang and Singh (2007) and Kao and Govindaraju (2007) modeled the dependence between peak storm intensity and depth, depth and duration and peak intensity and duration.

One of the most popular parametric families of copulas is the Cook-Johnson (1981) copula family, defined by

$$C^{CJ}(u, v) = [(1-u)^{\frac{1}{\beta}} + (1-v)^{\frac{1}{\beta}} - 1]^{-\beta},$$
$$\beta > 0. \quad (1)$$

Inverting this copula results in the joint function:

$$H(x, y) = [x + y - 1]^{-\beta},$$
$$x \geq 1, y \geq 1, \beta > 0, \quad (2)$$

where X and Y are identical type I Pareto distributions $F(x) = 1 - x^{-\beta}$ and $F(y) = 1 - y^{-\beta}$, respectively (Genest & Rivest, 1993).

This study extends the copula family of Cook-Johnson by considering a weighted function and examines values of dependence. In addition, the new family is described and examples of copulas taken in this family are

Hakim Bekrizadeh, Gholam Ali Parham and Mohamd Reza Zadkarmi are in the Department of Statistics. Email them at: H_bekri@yahoo.com, Parham_g@scu.ac.ir and Zadkarmi_m@scu.ac.ir.

provided. The associated Kendall's τ_k and tail dependence coefficient are also considered and an application of the generalized Cook-Johnson copula in analysis of probable modeling of a hot spring eruption is presented.

Weighted Cook-Johnson Copulas

A new family of generalized Cook-Johnson copula is proposed using a type II Pareto weighted distribution function. The weighted distributions occur when a random sample from an entire population of interest cannot be obtained (as in the tails) or when a random sample is not desired (as in the selection models).

Let X be a random variable with density $f(x)$ and $w(x)$ be a non-negative real value function such that $E[w(X)] < \infty$. The weighted random variable X , for example X^w , then has weighted probability density function (pdf) of

$$f_{X^w}(x) = \frac{w(x)f_X(x)}{E[w(X)]} \quad (3)$$

(Patil & Rao, 1978; Mahfoud & Patil, 1982).

If X is a type I Pareto distribution and $w(x) = x^{-\alpha}$, $\alpha > 0$, is a non-negative real valued function, then the weighted pdf of X is

$$f_{X^w}(x) = (\alpha + \beta)x^{-(\alpha+\beta)-1}, \quad x > 1, \alpha, \beta > 0, \quad (4)$$

and the survival weighted function is

$$F_{X^w}(x) = 1 - x^{-(\alpha+\beta)}, \quad x > 1, \alpha, \beta > 0 \quad (5)$$

Cook-Johnson Copula Theorem

Let (X, Y) be a type I joint Pareto distribution. Under the weight function $W(x, y) = [x + y - 1]^{-\alpha}$ for $\alpha > 0$, the weighted copula is

$$C_a^{CJ}(u, v) = [(1-u)^{-\frac{1}{\beta+\alpha}} + (1-v)^{-\frac{1}{\beta+\alpha}} - 1]^{-(\beta+\alpha)}, \quad 0 < u, v < 1, \alpha, \beta > 0, \quad (6)$$

and the type I Pareto joint weighted function is

$$H(x, y) = [x + y - 1]^{-(\beta+\alpha)}, \quad x, y > 1, \alpha, \beta > 0 \quad (7)$$

and X^w, Y^w are two identical weighted type I Pareto random variables with survival weighted functions

$$F_{X^w}(x) = 1 - x^{-(\alpha+\beta)}, \quad x > 1, \alpha, \beta > 0$$

and

$$F_{Y^w}(y) = 1 - y^{-(\alpha+\beta)}, \quad y > 1, \alpha, \beta > 0$$

respectively.

Cook-Johnson Copula Proof

For any pair of random variables (X, Y) , it can be shown from (6) that the joint pdf for (X, Y) is

$$f_{(X,Y)}(x, y) = \beta(\beta+1)(x+y-1)^{-(\beta+2)}, \quad x, y > 1, \beta > 0. \quad (8)$$

Under the weight function $W(x, y) = [x + y + 1]^{-\alpha}$ for $\alpha > 0$, a type I joint Pareto weighted density function of $(X, Y)^w$

$$h^w(x, y) = \frac{w(x, y)h(x, y)}{E[w(X, Y)]} \quad (9)$$

is

$$h^w(x, y) = (\alpha + \beta)(\alpha + \beta + 1)(x + y - 1)^{-(\alpha+\beta)+2}, \quad x, y > 1, \beta, \alpha > 0 \quad (10)$$

with marginal weighted density, such as (5), for X^w , Y^w and type I joint weighted function

$$H^w(x, y) = [x + y - 1]^{-(\beta + \alpha)}, \quad (11)$$

$$x, y > 1, \beta > \alpha > 0.$$

The marginal weighted functions F_{X^w} and G_{Y^w} are

$$F_{X^w}(x) = 1 - x^{-(\beta + \alpha)},$$

$$x > 1$$

and

$$G_{Y^w}(y) = 1 - y^{-(\beta + \alpha)},$$

$$y > 1$$

Inverting the weighted functions and employing the version of Corollary 2.3.7 (Nelson, 2006) yields the weighted copula

$$C_{\alpha}^{CJ}(u, v) = [(1-u)^{-\frac{1}{\beta + \alpha}} + (1-v)^{-\frac{1}{\beta + \alpha}} - 1]^{-(\beta + \alpha)},$$

$$0 < u, v < 1, \alpha, \beta > 0 \quad (12)$$

This relation may be called the weighted Cook-Johnson Copula. The copula density is given by:

$$c_{\alpha}^{CJ}(u, v) = \frac{\partial^2 C_{\alpha}^{CJ}(u, v)}{\partial u \partial v}$$

$$= \left[\frac{\beta + \alpha + 1}{\beta + \alpha} \left((1-u)^{-\frac{1}{\beta + \alpha} + 1} \right) \right] \left[(1-u)^{-\frac{1}{\beta + \alpha}} \right]^{-(\beta + \alpha) - 2}$$

$$\times \left[(1-v)^{-\frac{1}{\beta + \alpha} + 1} \right] \left[(1-v)^{-\frac{1}{\beta + \alpha}} - 1 \right]$$

Figure 1 illustrates the bivariate plots for the extended Cook-Johnson copula; the figure shows a surface bounded within a unit cube that is tied up along the two axes in the first quadrant. After placing different β 's and α 's, the graph of a generalized copula in terms of different β 's and α 's did not change. Figure 1 also shows upper correlations at the right in large quantities.

Figure 1: Bivariate Plots of the Extended Cook-Johnson Copulas

Cook-Johnson Copula Corollary

Under the assumption of the Cook-Johnson Copula theorem, the generalized weighted Cook-Johnson Copula for a d-dimensional is

$$C_{\alpha}^{CJ}(u_1, u_2, \dots, u_d) =$$

$$\left(\sum_{i=1}^d (1-u_i)^{-\frac{1}{\beta + \alpha}} - d + 1 \right)^{-(\beta + \alpha)},$$

and for a sub-dimensional it is

$$C_{\alpha}^{CJ}(u_{k+1}, u_{k+2}, \dots, u_d) =$$

$$\left(\sum_{i=k+1}^d (1-u_i)^{-\frac{1}{\beta + \alpha}} - (d - k - 1) \right)^{-(\beta + \alpha)}.$$

The generalized copula can be used to make bivariate distributions.

Generalized Weighted Bivariate Beta Distribution Example

If $X \sim \text{beta}(\theta, 1)$ and $Y \sim \text{beta}(\theta, 1)$, then $u = x^\theta$, $v = y^\theta$ and the generalized weighted bivariate beta distribution is

$$H(x, y) = [(1 - x^\theta)^{-\frac{1}{\alpha+\beta}} + (1 - y^\theta)^{-\frac{1}{\alpha+\beta}} - 1]^{-(\alpha+\beta)},$$

$$0 < x, y < 1, \alpha, \beta, \theta > 0.$$

Note that, when $\theta = 1$, the generalized weighted bivariate uniform distribution becomes:

$$H(x, y) = [(1 - x)^{-\frac{1}{\alpha+\beta}} + (1 - y)^{-\frac{1}{\alpha+\beta}} - 1]^{-(\alpha+\beta)},$$

$$0 < x, y < 1, \alpha, \beta > 0.$$

Generalized Weighted Bivariate Weibull Distribution Example

If $X \sim W(\theta, \lambda)$ and $Y \sim W(\theta, \lambda)$, then $u = 1 - e^{-\lambda x^\theta}$, $v = 1 - e^{-\lambda y^\theta}$ and the generalized weighted bivariate Weibull distribution is

$$H(x, y) = [(1 - e^{-\lambda x^\theta})^{-\frac{1}{\alpha+\beta}} + (1 - e^{-\lambda y^\theta})^{-\frac{1}{\alpha+\beta}} - 1]^{-(\alpha+\beta)},$$

$$0 < x, y, \alpha, \beta, \theta, \lambda.$$

It should be noted that, when $\theta = 1$, the generalized weighted bivariate exponential distribution is obtained as

$$H(x, y) = [(1 - e^{-\lambda x})^{-\frac{1}{\alpha+\beta}} + (1 - e^{-\lambda y})^{-\frac{1}{\alpha+\beta}} - 1]^{-(\alpha+\beta)},$$

$$0 < x, y, \alpha, \beta, \lambda;$$

And, when $\theta = 2$, by replacing $\frac{1}{2\lambda}$ as opposed to λ , the generalized weighted bivariate Rayleigh distribution is obtained as

$$H(x, y) = [(1 - e^{-\frac{x^2}{2\lambda}})^{-\frac{1}{\alpha+\beta}} + (1 - e^{-\frac{y^2}{2\lambda}})^{-\frac{1}{\alpha+\beta}} - 1]^{-(\alpha+\beta)},$$

$$0 < x, y, \alpha, \beta, \lambda.$$

Calculating Measures of Dependence

Copulas can be used in the study of dependence between random variables in many different ways. For a historical review of association measures and concepts of independence, see Gebremichael and Krajewski (2007), Genest and Rivest (1993) and Hutchinson and Lai (1990).

Kendall's τ_k

If X and Y are continuous random variables with copula C , then the population version of Kendall's tau for X and Y (denoted by τ_k) is given by

$$\tau_k = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1, \quad (13)$$

where C is the copula associated to (X, Y) . Thus, the Kendall's τ_k of the generalized Cook-Johnson copula is given by:

$$\tau_k = 1 - \frac{2}{2(\beta + \alpha) + 1}.$$

Tail Dependence

The concept of tail dependence relates to the amount of dependence in the upper-right quadrant tail or lower-left-quadrant tail of a bivariate distribution (Farlie, 1960). It is a concept that is relevant for the study of dependence between extreme values. Tail dependence between two continuous random variables X and Y is a copula property, hence, the amount of tail dependence is invariant under strictly increasing transformations of X and Y .

Definition 1

If a bivariate copula C is such that

$$L_U = \lim_{u \rightarrow 1} \frac{1 - 2u + C(u, u)}{1 - u} \quad (14)$$

exists, then C has upper tail dependence if $L_U \in (0, 1]$ and upper tail independence if $L_U = 0$. This measure is used extensively in

extreme value theory; it is the probability that one variable is extreme given that the other is extreme, that is, $L_U = P(U > u | V > u)$. Thus L_U may be viewed as a quantile dependent measure of dependence (Coles, Currie & Tawn, 1999).

Definition 2

The concept of lower tail dependence can be defined in a similar way. If the limit

$$L_L = \lim_{u \rightarrow 0} \frac{C(u, u)}{u} \quad (15)$$

exists, then C has lower tail dependence if $L_L \in (0, 1]$ and lower tail independence if $L_L = 0$. Similarly, lower tail dependence is defined as $L_L = P(U < u | V < u)$. For copulas without a simple closed form an alternative formula for L_L is more useful; thus, the upper tail dependence of the generalized Cook-Johnson copula using (14) is $L_U = 2 - 2^{-\frac{1}{\beta + \alpha}}$, and the lower tail dependence of the generalized Cook-Johnson copula using (15) is $L_L = 0$.

Application of Weighted Cook-Johnson Copula

There is a need for information in the analysis and management of risk for hot spring eruption, the most important of which are the frequency of time between two eruptions and the distance of eruptions. Data used in this study was collected in Yellowstone National Park in 1978 (Weisberg, 1985). Considering the high correlation of these two features, some tools must be used to determine the amount of relationship and impact that exists in the analysis of hot spring eruption; it is necessary to determine the joint distribution of the two features, time interval between two eruptions and distance of an eruption. Because the correlation between the two factors is 0.841 and the hypothesis of independence is not significant, the joint distribution of time interval between two eruptions and distance of an eruption is difficult to obtain via an estimate of marginal distribution. For this reason it

is necessary to estimate the marginal distribution of each of the two factors, time interval between two eruptions and distance of an eruption, and between the family of functions selected, the family having the conditions for modeling hot spring eruption.

In most studies in hydrology and geology, exponential, gamma and Weibull distributions are fitted to data so they are used herein and, for data of time interval between two eruptions, the exponential distribution with the parameter $\lambda_1 = 3.464152$ ($Sig. = 0.000$); for distance of an eruption, the exponential distribution with parameter $\lambda_2 = 71.13208$ ($Sig. = 0.000$) is a better fit. For estimating each parameter, a copula function is inserted and the obtained function is at a maximum based on the β parameter due to the complexity of the form of density function used for the family and time of plan; therefore, to estimate the β parameter the likelihood function logarithm is defined, for drawn β and by limiting the range obtain the likelihood function of maximum logarithm. All related calculations to distribution function, estimating parameter and likelihood function logarithm, are obtained using Maple software.

The parameter estimator for the Cook-Johnson copula function is $\beta = 14.435$ and the maximum value of the likelihood logarithm is -340.920279 . Given that sampling devices collect data with some restrictions based on the mechanism used to record observations, it is possible that the proportional is a non-negative function of their size because the observations have a weight distribution. Thus, for data regarding time interval between two eruptions, the type I Pareto distribution with parameters $\mu_1 = 192.35$ and $\sigma_1 = 726.09$ ($Sig. = 0.000$), for distance of an eruption with a type I Pareto distribution with parameters $\mu_2 = 208.01$ and $\sigma_2 = 8940.6$ ($Sig. = 0.000$), and by a defined weight function with parameters $\alpha = 5.623$ (see theorem), then the maximum estimated parameters for the weighted Cook-Johnson copula function are $\beta = 9.875$ and $\alpha = 5.623$.

The maximum value of the likelihood logarithm for the weighted Cook-Johnson copula

is -295.012384 . Better copula function is selected according to the method of likelihood maximum, thus, the weighted Cook-Johnson copula function provides more changes for observing samples related to estimated value of Cook-Johnson function. The weighted Cook-Johnson copula function with the parameters $\beta = 9.875$ and $\alpha = 5.623$ can then be used to determine the distribution combined with the time interval between two eruptions and distance of eruption. The estimated value of Kendall's tau is 0.937, the maximum likelihood estimate of the parameters of $\beta = 9.875$ and $\alpha = 5.623$, and for the fitted type I Pareto distribution estimates for the parameters are $\mu_1 = 192.35$, $\sigma_1 = 726.09$, $\mu_2 = 208.01$ and $\sigma_2 = 8940.6$.

In the weighted Cook-Johnson copula the distribution of time interval between two eruptions and distance of eruption is:

$$H(x, y) = \left[\left(\frac{192.35}{x} \right)^{-46.85} + \left(\frac{208.01}{y} \right)^{-576.89} - 1 \right]^{-17.498}$$

and the density is given by

$$h(x, y) = \left(1.06 \left(\frac{192.35}{x} \right)^{-679.24} \times \left(\frac{208.01}{y} \right)^{-8363.71} \right) \left[\left(\frac{192.35}{x} \right)^{-46.85} + \left(\frac{208.01}{y} \right)^{-576.89} - 1 \right]^{-17.498}$$

Using the distribution of an eruption, important information can be obtained regarding the eruption of hot spring, for example, the probability of an eruption of a hot spring with a time of 3.5 hours between two eruptions and a distance of one eruption of 42 meters is 0.0053. Using the copula function and a marginal distribution allows probabilities and other information about eruptions of hot spring and the relationship between the time intervals and distances of eruption to be obtained. Conditional distribution can also be determined by copula to provide a basis for the probabilities of altering factors against controlled changes. This may be useful in understanding and managing the

impacts of global warming on hot spring eruptions.

Conclusion

This article proposed a new family of copulas; the generalizing Cook-Johnson family generated by weighted distribution function, and obtained a generalized d -dimensional (multivariate) Cook-Johnson Copula. The bivariate weighted distributions and weighted marginal distribution were also generated, and the generalized bivariate distribution was obtained. The main feature of this family of copulas is to permit modeling of variables with high dependence. Moreover, it was shown that the generalizing Cook-Johnson Copula is a proper model for analyzing the problem of eruption of a hot spring. Bivariate distributions and copula functions for two parameters were analyzed. By using research methodology and the multivariate generalizing Cook-Johnson copula function, the issue of probable modeling of hot spring eruption with more variables can be studied.

References

- Coles, S., Currie, J., & Tawn, J. (1999). *Dependence measures for extreme value analyses*. Working paper: Department of Mathematics and Statistics, Lancaster University.
- Cook-Johnson, D. G. (1981). A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incident. *Biometrika*, 65, 141-151.
- De Michele, C., et al. (2005). Bivariate statistical approach to check adequacy of dam spillway. *Journal of Hydrologic Engineering*, 10(1), 50-57.
- Farlie, D. G. J. (1960). The performance of some correlation coefficients for a general bivariate distribution. *Biometrika*, 47, 307-323.
- Gebremichael, M., & Krajewski, W. F. (2007). Application of copulas to modeling temporal sampling errors in satellite-derived rainfall estimates. *Journal of Hydrologic Engineering*, 12(4), 404-408.
- Genest, C., & Rivest, L. (1993). Statistical Inference Procedures for bivariate Archimedean Copulas. *Journal of the American Statistical Association*, 88, 1034-1043.

- Genest, C., & Favre, A. C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrological Engineering*, **12**, 347–368.
- Hutchinson, T. P., & Lai, C. D. (1990). *Continuous bivariate distributions, emphasising applications*. Adelaide: Rumsby Scientific Publishing.
- Joe, H. (1997). *Multivariate models and dependence concepts*. London: Chapman & Hall.
- Kao, S. C., & Govindaraju, R. S. (2007). A bivariate frequency analysis of extreme rainfall with implications for design. *Journal of Geophysical Research, Atmospheres*.
- Mahfoud, M., & Patil, G. P. (1982). On weighted distributions. In *Statistics in theory and Practice: Essays in Honor of C. R. Rao*, G. Kallianpur, et al. (Eds.), 479-492. Amsterdam: North-Holland.
- Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd Ed. New York, NY: Springer-Verlag, Inc.
- Patil, G. P., & Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, **34**, 179-89.
- Salvadori, G., De Michele, C. (2004a). Analytical calculation of storm volume statistics involving Pareto-like intensity-duration marginals. *Geophysical Research Letters*, **31**, L04502.1-L04502.4.
- Salvadori, G., De Michele, C. (2004b). Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. *Water Resources Research*, **40**, W12511.1-W12511.17.
- Schweizer, B., & Sklar, A. (1983). *Probabilistic Metric Spaces*. Elsevier New York.
- Weisberg S. (1985). *Applied linear regression*. New York, NY: Wiley.
- Zhang, L., & Singh, V. P. (2007). Bivariate rainfall frequency distributions using Archimedean copulas. *Journal of Hydrology*, **332**(1-2), 93-109.

Multivariate Generalized Poisson Distribution for Inference on Selected Non-Communicable Diseases in Lagos State, Nigeria

Adewara Johnson Ademola Mbata Ugochukwu Ahamefula
University of Lagos,
Akoka, Lagos, Nigeria

Multivariate Generalized Poisson Distribution (MGPD) models are applied to make inferences regarding non-communicable diseases, diabetes, hypertension, stroke and ulcer in Lagos State, Nigeria. The generalized Poisson distribution is employed due to its usefulness in modeling count data in the presence of either over- or under- dispersion. Results show that the correlation between ulcer and stroke is not significant. Other pairwise comparisons of diseases are significant, thus implying that a patient who suffers from diabetes or stroke has a high propensity to also be hypertensive.

Key words: Multivariate generalized Poisson distribution, non-communicable disease, correlation, pairwise comparison.

Introduction

A multivariate generalized Poisson distribution (MGPD) is a discrete probability distribution that has the capacity to estimate the marginal mean and variance, which are univariate generalized Poisson distributions (GPD). MGPD allows for any form of correlation and can be used to describe count data with any type of dispersion. Several studies and applications have been conducted in count data modeling but fewer numbers of studies use MGPD models. Ordinary Poisson models have little ability to manage dispersion; in such cases, generalized Poisson distributions are used. According to Famoye, et al. (2011), a Poisson distribution satisfies the equi-dispersion property because its mean equals the variance, however, this property does not hold when the data is over-dispersed

(when the variance exceeds the mean) or under-dispersed (when the variance is less than the mean). When data is over- or under- dispersed, it is necessary to employ a probability model.

Among the alternatives to the Poisson distribution are the negative binomial distribution (NBD), the generalized Poisson distribution (GPD) and the Poisson log-normal distribution (PLD). The generalized Poisson distribution is relevant because it has the capacity to measure either over- or under-dispersion. Notable works in this area include Consul (1989), Vernic (1997, 1999, 2000), Tsiamyrtzis and Karlis (2004) and Famoye (2010). This article estimates the mean prevalence of selected non-communicable diseases using correlation. A monthly retrospective investigation and routine check of hospital records of patients was used to count the number of patients living with any of these health problems at the General Hospital, Lagos, Nigeria from January 2005 to March 2012.

Adewara Johnson Ademola is a Lecturer at Distance Learning Institute, University of Lagos. He earned his Ph.D. in statistics. Email him at: adewaraja@yahoo.com. Mbata Ugochukwu Ahamefula is an Assistant Lecturer in the Department of Mathematics. He is a Ph.D. student in statistics. Email him at: mbataugochukwu@yahoo.com.

Non-Communicable Diseases

Onwasigwe (2010) defined non-communicable diseases (NCDs) as diseases which are not contagious, that is, they cannot be transmitted from one person to another. Non-communicable diseases are chronic conditions that do not result from an acute infectious process but cause death, dysfunction or

impairment in quality of life. NCDs typically develop over a relatively long time without causing symptoms but after the disease manifests there may be a period of prolonged impaired health.

The World Health Organization (WHO) in (2011a) released the first *Global Status Report on Non-Communicable Diseases* which outlines the statistics, evidence and experiences needed for a more forceful response to the growing threat posed by NCDs. WHO (2011b) further provided an overview of each country's profile, the number, rates and causes of deaths from NCDs, the prevalence of selected risk factors, trends in metabolic risk factors in each country and information describing current prevention and control of NCDs.

Non-communicable diseases constitute a leading cause of functionary impairment and death globally. Nigeria loses about 400 million dollars yearly in national income from premature deaths from diseases such as diabetes mellitus, hypertension, cancer, renal failure and stroke and the economic cost from premature deaths due to NCDs is expected to rise to 8 billion dollars (Ogbebo, 2011). The WHO (2011b) report of year 2011 put mortality from NCDs at 59.8% of the global mortality and NCDs were estimated to account for 27% of all deaths in Nigeria. Due to the growing problem with these diseases in Nigeria among both young and old people, this study examines the mean prevalence and correlation of diabetes, hypertension, stroke and ulcer.

Multivariate Generalized Poisson Distribution

Famoye, et al. (2011) defined a new multivariate (d -variate) generalized Poisson distribution (MGPD) as a product of generalized Poisson marginals with a multiplicative factor. The probability function of the MGPD is

$$P(y_1, \dots, y_d) = \prod_{t=1}^d \left[\frac{\theta_t^{y_t} (1 + \alpha_t y_t)^{y_t-1}}{y_t!} \exp[-\theta(1 + \alpha_t y_t)] \right] \times \left[1 + \sum_{t < v}^d \lambda_{tv} (e^{-y_t} - c_t) (e^{-y_v} - c_v) \right] \quad (1)$$

where $c_t = E(e^{-Y_t}) = \exp[\theta_t(s-1)]$ with $\ln s_t - \alpha_t \theta_t (s_t - 1) + 1 = 0$;

$$c_{tt} = E(Y_t e^{-Y_t}) = \theta_t (1 - \alpha_t \theta_t s_t)^{-1} e^{\theta_t(1-\alpha_t)(s_t-1)-1}$$

with $\ln s_t - \alpha_t \theta_t (s_t - 1) + 1 = 0$ and $y_1, \dots, y_d = 0, 1, 2, \dots$.

The variate y_t exhibits some dispersion when $\alpha_t \neq 0$. It is under-dispersed when $\alpha_t < 0$ and is over-dispersed when $\alpha_t > 0$. The variates y_t and y_v are correlated when $\lambda_{tv} \neq 0$. The pair is negatively (or positively) correlated $\lambda_{tv} < 0$ as (or $\lambda_{tv} > 0$). According to Famoye (2011), the d marginal means and d marginal variances of the MGPD are:

$$\mu_t = \theta_t (1 - \alpha_t \theta_t)^{-1}, \quad t = 1, 2, \dots, d \quad (2)$$

and

$$\mu_t = \theta_t (1 - \alpha_t \theta_t)^{-3}, \quad t = 1, 2, \dots, d \quad (3)$$

respectively. The $d(d-1)/2$ covariances between any two variates are

$$\sigma_{tv} = \lambda_{tv} (c_{tt} - c_t \mu_t) (c_{vv} - c_v \mu_v), \quad t, v = 1, 2, \dots, d, \text{ and } t < v. \quad (4)$$

Using the covariance σ_{tv} between the variables Y_t and Y_v in equation (4), the correlation coefficient between Y_t and Y_v is

$$\rho_{tv} = \sigma_{tv} / (\sigma_t \sigma_v) = \lambda_{tv} (c_{tt} - c_t \mu_t) (c_{vv} - c_v \mu_v) / (\sigma_t \sigma_v). \quad (5)$$

Thus, the parameter λ_{tv} can be written in terms of the correlation coefficient ρ_{tv} . The correlation coefficient can be positive, zero or negative depending on the value of λ_{tv} . The parameter

λ_{tv} satisfies $|\lambda_{tv}| \leq 1 / [(1 - c_t)(1 - c_v)]$. Using this result, the correlation coefficient satisfies the condition:

$$|\rho_{tv}| \leq \frac{(c_{tt} - c_t \mu_t)(c_{vv} - c_v \mu_v)}{[\sigma_t \sigma_v (1 - c_t)(1 - c_v)]}.$$

Parameter Estimation

Assuming n independent vectors $(y_{i1}, y_{i2} \dots y_{in})$, where the i^{th} vector has the MGPD in (1). The sample mean and sample variance are denoted by $\bar{y}_t = \sum_{i=1}^n y_{it} / n$ and $s_t^2 = \sum_{i=1}^n (y_{it} - \bar{y}_t)^2 / (n-1)$ where $(t = 1, 2, \dots, d)$ respectively. The sample covariance between the variables Y_t and Y_v is denoted by

$$s_{tv} = \sum_{i=1}^n (y_{it} - \bar{y}_t)(y_{iv} - \bar{y}_v) / (n-1)$$

Equating these sample moments to the corresponding population moments, the moment estimates for the MGPD are given by

$$\begin{aligned} \tilde{\theta}_t &= \sqrt{\bar{y}_t^3 / s_t^2}, \\ t &= 1, 2, \dots, d \end{aligned} \quad (6)$$

$$\begin{aligned} \tilde{\alpha}_t &= \tilde{\theta}_t^{-1} - \bar{y}_t^{-1}, \\ t &= 1, 2, \dots, d \end{aligned} \quad (7)$$

$$\tilde{\lambda}_{tv} = s_{tv} (\tilde{c}_{tt} - \tilde{c}_t \bar{y}_t) (\tilde{c}_{vv} - \tilde{c}_v \bar{y}_v)^{-1} \quad (8)$$

$$\frac{\tilde{\alpha}_t}{\tilde{\theta}_t} \times 100 = \text{Coefficient of Dispersion}, \quad (9)$$

where $\tilde{c}_t$ and $\tilde{c}_{tt}$ are the estimated values of c_t and c_{tt} (where $t, v = 1, 2, \dots, d$ and $t < v$). In general, for d -variates MGPD, equations (6)-(8) provide a total of $2d + d(d-1)/2$ equations, which are solved simultaneously to obtain the moment estimates. For the log-likelihood function and estimation of the parameters, see Famoye (2010) and Famoye, et al. (2011).

The log-likelihood function, $\text{Log } L = \log L(\underline{\theta}, \underline{\alpha}, \underline{\lambda}; y)$, for the MGPD is:

$$\log L = \sum_{i=1}^n \left\{ \sum_{t=1}^d \left[y_{it} \log \theta_t + (y_{it} - 1) \log (1 + \alpha_t y_{it}) \right] - \theta_t (1 - \alpha_t y_{it}) - \log (y_{it}!) \right. \\ \left. + \left[1 + \sum_{t < v}^d \lambda_{tv} (e^{-y_{it}} - c_t) (e^{-y_{iv}} - c_v) \right] \right\} \quad (10)$$

The log-likelihood in (10) is maximized over the parameters θ_t , α_t ($t = 1, 2, \dots, d$), and λ_{tv} ($t, v = 1, 2, \dots, d$ and $t < v$). Famoye, et al. (2011) concluded that a measure of goodness of fit for the MGPD may be based on the log-likelihood statistic given in (10).

Methodology

The data for this research was collected from Island Hospital, Lagos. The numbers of patients suffering from diabetes, hypertension, stroke and/or ulcer from January 2005 to March 2012 were counted. The total number of patients observed suffering from diabetes, hypertension, stroke and ulcer were: 7,898, 10,055, 8,565 and 5,604 respectively. These figures sum to 32,122 out of 61,786 patients observed, constituting 51.99%. The parameters $\bar{y}_t$ and s_t^2 were estimated using Excel and R statistical packages; descriptive statistics and correlations ρ_{tv} between the pairs of the diseases were also estimated.

Test for Constant Dispersion Parameter

A test of dispersion was conducted of the parameters for MGPD. The test was given as $\alpha_t \neq 0$ ($t = 1, 2, \dots, d$), with a test hypothesis for constant dispersion:

$$\begin{aligned} H_0: \alpha_1 &= \alpha_2 = \dots = \alpha_d \\ \text{vs} \\ H_1: \alpha_1 &\neq \alpha_2 \neq \dots \neq \alpha_d \end{aligned} \quad (11)$$

Let L_{con} be the likelihood function when H_0 is true and let L_a be the likelihood function when H_0 is false. The test statistic is given by $\chi_{con} = -2\log(L_{con} / L_a)$, which is approximately a Chi-square with $d - 1$ degrees of freedom (Famoye, et al., 2011).

Results

Data was analyzed using the MGPD model to obtain descriptive statistics and to determine the parameters $\tilde{\theta}_t$, $\tilde{a}_t$, $\tilde{\mu}_t$, and $\tilde{\sigma}_t^2$. The correlation between the variables under consideration (see Table 1) shows that ulcer and stroke are not significant, however, the remaining paired variables are significant. This indicates that a patient suffering from any of these diseases also has a high degree of risk for the other diseases. For example, a diabetic patient may also be expected to be hypertensive. However, inferences regarding the selected non communicable diseases were based on the fitted MGPD model (see Table 2).

As shown in Table 2, the estimates for the dispersion parameter $\tilde{a}_t$ which is greater than zero are 0.176, 0.162, 0.219 and 0.112 indicating an over-dispersion in the data collected for each disease. Moreover, the result of the coefficient of dispersion (CD) in percentage also supports a constant dispersion. Testing the hypothesis on constant dispersion of parameters; a Chi-square value of 7.34 on 3 degrees for freedom is obtained with a p -value of 0.3713. Result show that dispersion exists and that the dispersion parameters are constant. The parameter estimates from MGPD with their corresponding standard errors are presented in Table 2. It can be observed that the standard deviation from the MGPD and standard deviations from the sample information are approximately equal. Furthermore, it may be inferred that the mean prevalence of diabetic, hypertension, stroke and ulcer patients per month are approximately 91, 115, 99 and 64, respectively. In addition, the estimates are significant using the confidence interval provided by the estimate and standard error.

Table 1: Correlation and Summary of Descriptive Statistics

Disease	Diabetic	Hypertension	Stroke	Ulcer	$\bar{y}_t$	s_t^2	SD
Diabetic	1				90.782	26265.661	162.067
Hypertension	0.772*	1			115.575	45120.922	212.417
Stroke	0.639*	0.612*	1		98.448	50604.320	224.954
Ulcer	0.286*	0.388*	0.046	1	64.414	4378.222	66.168

*Correlation is significant at $\alpha = 0.01$ level

Conclusion

Count data was modeled using MGPD, which allows a determination regarding whether data is over- or under- dispersed. This is an advantage of the MGPD model over other discrete probability models such as negative binomial distribution (NBD), Poisson distribution (PD), Poisson log-normal distribution (PLD), and multivariate Poisson distribution (MPD). Out of 61,786 patients observed, 32,122 suffered from a non-communicable disease, constituting about 51.99%. This figure somewhat supports the statistical evidence highlighted by WHO (2011a, 2011b) regarding non-communicable diseases. Results from this investigation show that there is a high correlation between the pairs of the diseases diabetes, hypertension and stroke. Thus, it may be stated that a patient who suffers from diabetes or stroke is also likely to be hypertensive. Hence, continued study is highly recommended to investigate the threats posed by non-communicable diseases globally.

Acknowledgements

The authors express their profound gratitude to the entire staff and management of General Hospital, Lagos Island, Lagos, Nigeria, for their wonderful support during data collection.

References

- Consul, P. C. (1989). *Generalized Poisson distributions: Properties and applications*. New York: Marcel Dekker, Inc.
- Famoye, F., Okafor, R., & Adamu, M. (2011). A multivariate generalized Poisson distribution. *Journal of Statistical Theory and Applications*, 10(3), 519-531.
- Famoye, F., & Singh, K. P. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science*, 4, 117-130.
- Famoye, F. (2010). A new bivariate generalized Poisson distribution. *Statistica Neerlandica*, 64, 112-124.

Table 2: Parameter Estimates for Non Communicable Diseases Studied

Disease	$\tilde{\theta}_i$	$\tilde{a}_i$	$\tilde{\mu}_i$	$\tilde{\sigma}_i^2$	SD	Estimate ± Standard Error	CD(%)
Diabetic	5.337	0.176	90.762	26249.523	162.017	90.762 ± 17.370 *	3.298
Hypertension	5.849	0.162	115.451	44980.595	212.086	115.451 ± 22.738 *	2.770
Stroke	4.360	0.219	98.678	50546.837	224.826	98.678 ± 24.104 *	5.023
Ulcer	7.813	0.112	64.412	4377.858	66.165	64.412 ± 7.094*	1.434

*Significant at $\alpha = 0.05$ level; CD, coefficient of dispersion

Ogbebo, W. (2011). *Non-communicable diseases to kill 5 million Nigerians*. Excerpt on Nigeria Tribune by Health Reform Foundation of Nigeria (HERFON).

Onwasigwe, C. (2010). *Disease transition in sub-Saharan Africa: The case of non-communicable diseases in Nigeria*. University of Nigeria, Nsukka: Community Medicine & Epidemiology.

Sheth, R. K. (1998). The generalized Poisson distributions and a model of clustering from Poisson initial conditions. *Monthly Notices Royal Astronomical Society*, 299(1), 207-217.

Tsiamyrtzis, P., & Karlis, D. (2004). Strategies for efficient computation of multivariate Poisson probabilities. *Communications in Statistics Simulation and Computation*, 33(2), 271-292.

Vernic, R. (1997). On the bivariate generalized Poisson distribution. *ASTIN Bulletin*, 27, 23-31.

Vernic, R. (1999). Recursive evaluation of some bivariate compound distributions. *ASTIN Bulletin*, 29, 315-325.

Vernic, R. (2000). On the multivariate generalization of the generalized Poisson distribution. *ASTIN Bulletin*, 30, 57-67.

World Health Organization. (2011a). *World population prospects – the 2010 revision*. New York, NY: United Nations Population Division.

World Health Organization. (2011b). *Non-communicable diseases country profiles 2011*. Geneva, Switzerland: WHO Press.

World Health Organization. (2009). *World development indicators*. Washington, DC: International Bank for Reconstruction and Development/The World Bank.

Brief Report Posterior Estimates of Poisson Distribution using R Software

Raja Sultan S. P. Ahmad
University of Kashmir,
Srinagar, J & K, India

The Bayesian estimation of unknown parameter of the Poisson distribution is examined under different priors. The posterior distributions for the unknown parameter of the Poisson distribution are derived using the following priors: uniform, Jeffrey's, Gamma distribution, Gamma-Chi-square distribution, Gamma-exponential distribution and Chi-square-exponential distribution. Numerical and graphical illustrations of the posterior densities of the parameters of interest were conducted using R Software.

Key words: Poisson distribution, prior distribution, posterior distribution, R software.

Introduction

The Poisson distribution was discovered in 1837 by French mathematician and physicist Simon Dennis Poisson (1781-1840). The Poisson distribution has many practical applications; one major area of application is in epidemiology, the study of disease incidence. The Poisson distribution arises naturally as a useful model for analyzing counts of rare events, such as the number of radioactive emissions in a fixed time interval or the number of bacteria in a given test sample.

The probability mass function (pmf) of the Poisson distribution of a random variable Y having Parameter λ is given by

$$f(y) = \frac{e^{-\lambda} \lambda^y}{y!} \quad (1)$$

$y = 0, 1, 2, \dots$ and $\lambda > 0$.

Haight's (1967) handbook provides many references regarding the Poisson distribution's applications in a variety of fields. Howlader and Balasooriya (2003), Bolstad (2007), Saleem and Aslam (2008), Tahir and Zawar (2008), Aslam and Haq (2009) and others have also discussed the distribution with applications.

Posterior Distribution of Unknown Parameter λ Using Uniform Prior

If $y_1, y_2, \dots, y_n$ are iid observations from an (1), then the likelihood function is

$$L(\lambda) = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!} \quad (2)$$

Laplace (1774, 1812) found that it worked exceptionally well to always select the prior for λ to be constant. Consider the uniform prior

$$P(\lambda) \propto 1, \quad (3)$$

$$0 < \lambda < \infty;$$

the posterior distribution $P(\lambda | Y)$ of the parameter λ is found using (2) and (3) as

Raja Sultan is Research Scholar in the Department of Statistics. Email him at: hamzasultan18@yahoo.com. S. P. Ahmad is an Assistant Professor in the Department of Statistics. Email him at: sprvz@yahoo.com.

$$P(\lambda | Y) \propto P(\lambda)$$

$$\Rightarrow P(\lambda | Y) = \frac{n^{\sum_{i=1}^n y_i + 1}}{\Gamma(\sum_{i=1}^n y_i + 1)} e^{-n\lambda} \lambda^{\sum_{i=1}^n y_i}, \quad (4)$$

which is the density function of a gamma distribution with parameters $\left(n, \sum_{i=1}^n y_i + 1\right)$.

Posterior Distribution of Unknown Parameter λ Using Jeffrey's Prior

The Jeffrey's prior (see Berger, 1985) for the parameter λ having distribution (1) is

$$P(\lambda) \propto \lambda^{-\frac{1}{2}} \quad (5)$$

and the posterior distribution $P(\lambda | Y)$ of the parameter λ is found using (2) and (5) as

$$P(\lambda | Y) = \frac{n^{\sum_{i=1}^n y_i + \frac{1}{2}}}{\Gamma(\sum_{i=1}^n y_i + \frac{1}{2})} e^{-n\lambda} \lambda^{\sum_{i=1}^n y_i - \frac{1}{2}} \quad (6)$$

which is the density function of a gamma distribution with parameters $\left(n, \sum_{i=1}^n y_i + \frac{1}{2}\right)$.

Posterior Distribution of Unknown Parameter λ Using Gamma Distribution as Prior

The single prior distribution of λ is a gamma distribution with hyper parameters a and b is

$$P(\lambda) = \frac{a^b e^{-a\lambda} \lambda^{b-1}}{\Gamma(b)} \quad (7)$$

$a > 0, b > 0, \lambda > 0.$

The posterior distribution $P(\lambda | Y)$ of parameter λ is found using (2) and (7) as

$$P(\lambda | Y) = \frac{(a+n)^{b+\sum_{i=1}^n y_i}}{\Gamma(b+\sum_{i=1}^n y_i)} e^{-\lambda(a+n)} \lambda^{(b+\sum_{i=1}^n y_i)-1}, \quad (8)$$

which is the density function of a gamma distribution with parameters

$$\left(a+n, b+\sum_{i=1}^n y_i\right).$$

Posterior Distribution of Unknown Parameter λ Using Gamma-Chi-Square Prior

Assume that the prior distribution of λ is a gamma distribution with hyper parameters a_1 and b_1 as

$$P_1(\lambda) = \frac{b_1^{a_1} e^{-\lambda b_1} \lambda^{a_1-1}}{\Gamma(a_1)}, \quad (9)$$

$a_1 > 0, b_1 > 0, \lambda > 0.$

The second prior assumed is a Chi-square distribution with hyper parameter c_1 given by

$$P_2(\lambda) = \frac{\left(\frac{1}{2}\right)^{\frac{c_1}{2}} e^{-\frac{\lambda}{2}} \lambda^{\frac{c_1}{2}-1}}{\Gamma\left(\frac{c_1}{2}\right)}, \quad (10)$$

$c_1 > 0, \lambda > 0.$

A double prior is defined for λ by combining the two priors in (9) and (10) as:

$$P(\lambda) \propto P_1(\lambda) P_2(\lambda)$$

$$\Rightarrow P(\lambda) \propto \left(\lambda^{a_1-1} e^{-\lambda(b_1+\frac{1}{2})} \lambda^{\frac{c_1}{2}-1} \right). \quad (11)$$

The posterior distribution $P(\lambda | Y)$ of parameter λ is found by using (2) and (11) as

$$P(\lambda | Y) = \frac{\left(n + b_1 + \frac{1}{2}\right)^{a_1 + \frac{c_1}{2} + \sum_{i=1}^n y_i - 1}}{\Gamma\left(a_1 + \frac{c_1}{2} + \sum_{i=1}^n y_i - 1\right)} e^{-\lambda(n + b_1 + \frac{1}{2})} \lambda^{(a_1 + \frac{c_1}{2} + \sum_{i=1}^n y_i - 1) - 1},$$

which is the density function of gamma distribution with parameters

$$\left(n + b_1 + \frac{1}{2}, a_1 + \frac{c_1}{2} + \sum_{i=1}^n y_i - 1\right).$$

Posterior Distribution of Unknown Parameter λ Using Gamma-Exponential Prior

The double prior for λ is defined to be a gamma distribution with hyper parameters (a_2 , b_2) and an exponential distribution with hyper parameter c_2 as

$$P(\lambda) \propto \lambda^{a_2 - 1} e^{-\lambda(b_2 + c_2)}, \quad \lambda > 0. \quad (12)$$

The posterior distribution $P(\lambda | Y)$ of parameter λ is given by using (2) and (12) as

$$P(\lambda | Y) = \frac{(n + b_2 + c_2)^{a_2 + \sum_{i=1}^n y_i}}{\Gamma\left(a_2 + \sum_{i=1}^n y_i\right)} e^{-\lambda(n + b_2 + c_2)} \lambda^{(a_2 + \sum_{i=1}^n y_i) - 1},$$

which is the density function of a gamma distribution with parameters

$$\left(n + b_2 + c_2, a_2 + \sum_{i=1}^n y_i\right).$$

Posterior Distribution of Unknown Parameter λ Using Chi-Square and Exponential Prior

The double prior for λ is defined to be a Chi-square distribution with hyper parameter a_3 and an exponential distribution with hyper parameter c_3 as

$$P(\lambda) \propto \lambda^{\frac{a_3}{2} - 1} e^{-\lambda(c_3 + \frac{1}{2})}, \quad \lambda > 0. \quad (13)$$

The posterior distribution $P(\lambda | Y)$ of parameter λ is found by using (2) and (13) as

$$P(\lambda | Y) = \frac{\left(n + c_3 + \frac{1}{2}\right)^{\frac{a_3}{2} + \sum_{i=1}^n y_i}}{\Gamma\left(\frac{a_3}{2} + \sum_{i=1}^n y_i\right)} e^{-\lambda(n + c_3 + \frac{1}{2})} \lambda^{(\frac{a_3}{2} + \sum_{i=1}^n y_i) - 1},$$

which is the density function of gamma distribution with parameters

$$\left(n + c_3 + \frac{1}{2}, \frac{a_3}{2} + \sum_{i=1}^n y_i\right).$$

Methodology

Numerical Illustration

The Poisson distribution provides a realistic model for many random phenomena. Because the values of a Poisson random variable are non-negative integers, any random phenomena for which a count is of interest is a candidate for modeling by assuming a Poisson distribution. The numerical and graphical illustration of posterior densities of the parameters of interest conveys a convincing and comprehensive picture of Bayesian data analysis.

Several programs were developed to calculate posterior densities of the Poisson distribution under various priors in R Software for this study. These programs illustrate the strength of Bayesian methods in practical situations. Data analyzed for this study is from Hoff (2009): During the 1990s the General Social Survey gathered data on educational attainment and number of children among 155 women who were 40 years of age at the time of their participation in the survey. Let $Y_{1,1}, Y_{2,1}, \dots, Y_{n1,1}$ denote the number of children for n_1 women without college degrees and $Y_{1,2}, Y_{2,2}, \dots, Y_{n2,2}$ denote the number of children

for n_2 women with college degrees. The group sums and means are:

Group I, Less than bachelors:

$$\begin{aligned} n_1 &= 111, \\ \sum_{i=1}^{n_1} Y_{i,1} &= 217, \\ \bar{Y}_1 &= 1.95 \end{aligned}$$

Group II, Bachelors or higher:

$$\begin{aligned} n_2 &= 44, \\ \sum_{i=1}^{n_2} Y_{i,2} &= 66, \\ \bar{Y}_2 &= 1.50 \end{aligned}$$

the posterior mean and posterior variance of parameter λ are presented in Table 1 for Groups I and II under different types of priors. Graphical displays of posterior densities for λ under different priors are shown in Figures 1 and 2.

Figure 1: Posterior Distribution of Mean Birth Rate for Group I

Figure 2: Posterior Distribution of Mean Birth Rate for Group II

Results

Table 2 shows that the posterior mean for Group I (less than bachelors) under all assumed priors is higher than that of Group II (bachelors or higher). The posterior variance for Group I (less than bachelors) is less than the posterior variance of Group II (bachelors or higher) under all assumed priors. The posterior variance under all the assumed priors is calculated assuming the values of all hyper parameters are 2. Table 2 also shows that the posterior variance under the double prior Gamma-Exponential distribution is less compared to other assumed priors; thus, this prior is more efficient compared to other priors and the lower variation in posterior distribution assists in more precise Bayesian estimates of the true unknown parameter λ of a Poisson distribution.

References

- Aslam, M., & Haq, A. (2009). On the prior selection for variance of normal distribution. *Interstat*, #005.
- Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*, 2nd Edition. New York, NY: Springer-Verlag.

POSTERIOR ESTIMATES OF POISSON DISTRIBUTION USING R SOFTWARE

Table 1: Posterior Mean and Posterior Variance of Parameter λ of a Poisson Distribution with Different Priors

Type of Prior	Prior Distribution	Posterior Mean	Posterior Variance
Uniform Prior	1	$\frac{\sum_{i=1}^n y_i + 1}{n}$	$\frac{\sum_{i=1}^n y_i + 1}{n^2}$
Jeffrey's Prior	$\lambda^{-\frac{1}{2}}$	$\frac{\sum_{i=1}^n y_i + \frac{1}{2}}{n}$	$\frac{\sum_{i=1}^n y_i + \frac{1}{2}}{n^2}$
Gamma Distribution	$e^{-a\lambda} \lambda^{b-1}$	$\frac{b + \sum_{i=1}^n y_i}{a + n}$	$\frac{b + \sum_{i=1}^n y_i}{(a + n)^2}$
Gamma-Chi-Square Distribution	$\lambda^{a_1-1} e^{-\lambda(b_1+\frac{1}{2})} \lambda^{\frac{c_1}{2}-1}$	$\frac{a_1 + \frac{c_1}{2} + \sum_{i=1}^n y_i - 1}{n + b_1 + \frac{1}{2}}$	$\frac{a_1 + \frac{c_1}{2} + \sum_{i=1}^n y_i - 1}{\left(n + b_1 + \frac{1}{2}\right)^2}$
Gamma-Exponential Distribution	$\lambda^{a_2-1} e^{-\lambda(b_2+c_2)}$	$\frac{a_2 + \sum_{i=1}^n y_i}{n + b_2 + c_2}$	$\frac{a_2 + \sum_{i=1}^n y_i}{(n + b_2 + c_2)^2}$
Chi-Square-Exponential Distribution	$\lambda^{\frac{a_3}{2}-1} e^{-\lambda(c_3+\frac{1}{2})}$	$\frac{\frac{a_3}{2} + \sum_{i=1}^n y_i}{n + c_3 + \frac{1}{2}}$	$\frac{\frac{a_3}{2} + \sum_{i=1}^n y_i}{\left(n + c_3 + \frac{1}{2}\right)^2}$

Table 2: Posterior Mean and Posterior Variance of a Poisson Distribution with Different Priors

Type Of Prior	Less Than Bachelor's		Bachelors Or Higher	
	Posterior Mean	Posterior Variance	Posterior Mean	Posterior Variance
Uniform Prior	1.963964	0.01769337	1.522727	0.03460744
Jeffrey's Prior	1.959459	0.01765279	1.511364	0.03434917
Gamma Distribution	1.938053	0.01715091	1.478261	0.03213611
Gamma-Chi-Square Distribution	1.929515	0.01700014	1.462366	0.03144872
Gamma-Exponential Distribution	1.904348	0.01655955	1.416667	0.02951389
Chi-Square-Exponential Distribution	1.920705	0.01692251	1.440860	0.03098624

Bolstad, W. M. (2007). *Introduction to Bayesian statistics*. New York, NY: John Wiley & Sons.

Haight, F. A. (1967). *Handbook of the Poisson distribution*. New York, NY: John Wiley & Sons.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York, NY: Springer.

Howlader, H. A., & Balasooriya, U. (2003). Bayesian estimation of the distribution function of the Poisson model. *Biometrical Journal*, 45(7), 901-912.

Laplace, P. S. (1774, trans 1986). Memoire sur la probabilite des causes par les evenemens. Translation by S. Stigler. Memoir on the probability of causes of events. *Statistical Science*, 1, 364-378.

Laplace, P. (1812). *Theorie Analytique des Probabilite's*. Paris: Courcier.

Saleem, M., & Aslam, M. (2008). On the prior selection for the mixture of Rayleigh distribution using predictive intervals. *Pakistan Journal of Statistics*, 24(1), 21-35.

Tahir, M., & Zavar, H. (2008). Comparison of non-informative priors for number of defects (Poisson) model. *Interstat*, #002.

Instructions For Authors

Follow these guidelines when submitting a manuscript:

1. *JMASM* uses a modified American Psychological Association style guideline.
2. Submissions are accepted via e-mail only. Send them to the Editorial Assistant at ea@jmasm.com. Provide name, affiliation, address, e-mail address, and 30 word biographical statements for all authors in the body of the email message.
3. There should be no material identifying authorship except on the title page. A statement should be included in the body of the e-mail that, where applicable, indicating proper human subjects protocols were followed, including informed consent. A statement should be included in the body of the e-mail indicating the manuscript is not under consideration at another journal.
4. Provide the manuscript as an external e-mail attachment in MS Word for the PC format only. (Wordperfect and .rtf formats may be acceptable - please inquire.) Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are **NOT** acceptable for manuscript submission.
5. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times New Roman font.
6. Create tables without boxes or vertical lines. Place tables, figures, and graphs “in-line”, not at the end of the manuscript. Figures may be in .jpg, .tif, .png, and other formats readable by Adobe Photoshop.
7. The manuscript should contain an Abstract with a 50 word maximum, following by a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left-justified, indent optional. Do not number headings or subheadings.
8. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline.
9. Do not use underlining in the manuscript. Do not use bold, except for (a) matrices, or (b) emphasis within a table, figure, or graph. Do not number references. Do not use footnotes or endnotes.
10. In the References section, do not put quotation marks around titles of articles or books. Capitalize only the first letter of books. Italicize journal or book titles, and volume numbers. Use “&” instead of “and” in multiple author listings.
11. *Suggestions for style*: Instead of “I drew a sample of 40” write “A sample of 40 was selected”. Use “although” instead of “while”, unless the meaning is “at the same time”. Use “because” instead of “since”, unless the meaning is “after”. Instead of “Smith (1990) notes” write “Smith (1990) noted”. Do not strike spacebar twice after a period.

Print Subscriptions

Print subscriptions including postage for professionals are US \$95 per year; for graduate students are US \$47.50 per year; and for libraries, universities and corporations are US \$195 per year. Subscribers outside of the US and Canada pay a US \$10 surcharge for additional postage. Online access is currently free at <http://www.jmasm.com/>. Mail subscription requests with remittances to JMASM, P. O. Box 48023, Oak Park, MI, 48237. Email journal correspondence, other than manuscript submissions, to ea@jmasm.com.

Notice To Advertisers

Send requests for advertising information to ea@jmasm.com.