ISSN 1927-7032 (Print)
ISSN 1927-7040 (Online)

International Journal of Statistics and Probability

Vol. 7, No. 4 July 2018

CANADIAN CENTER OF SCIENCE AND EDUCATION

INTERNATIONAL JOURNAL OF STATISTICS AND PROBABILITY

An International Peer-reviewed and Open Access Journal for Statistics and Probability

International Journal of Statistics and Probability (ISSN: 1927-7032; E-ISSN: 1927-7040) is an open-access, international, double-blind peer-reviewed journal published by the Canadian Center of Science and Education. This journal, published **bimonthly** (January, March, May, July, September and November) in both **print and online versions**, keeps readers up-to-date with the latest developments in all areas of statistics and probability.

The scopes of the journal:

- Computational statistics
- Design of experiments
- Sample survey
- Statistical modelling
- Statistical theory
- Probability theory

The journal is included in:

- BASE
- Google Scholar
- JournalTOCs
- LOCKSS
- SHERPA/RoMEO
- Ulrich's

Submission Policy

Submission of an article implies that the work described has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the authorities responsible where the work was carried out. However, we accept submissions that have previously appeared on preprint servers (for example: arXiv, bioRxiv, Nature Precedings, Philica, Social Science Research Network, and Vixra); have previously been presented at conferences; or have previously appeared in other "non-journal" venues (for example: blogs or posters). Authors are responsible for updating the archived preprint with the journal reference (including DOI) and a link to the published articles on the appropriate journal website upon publication.

iThenticate Professional Plagiarism Prevention

The publisher and journals have a zero-tolerance plagiarism policy. We check the issue using two methods: a plagiarism prevention tool (iThenticate) and a reviewer check. All submissions will be checked by iThenticate before being sent to reviewers.

creative commons

We insist a rigorous viewpoint on the self-plagiarism. The self-plagiarism is plagiarism, as it fails to contribute to the research and science.

IJSP accepts both Online and Email submission. The online system makes readers to submit and track the status of their manuscripts conveniently. For any questions, please contact ijsp@ccsenet.com.

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Editorial Team

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Contents

Explaining Lord's Paradox in Introductory Statistical Theory Courses 1
Steven B. Kim

The Application of Text Mining Algorithms In Summarizing Trends in Anti-Epileptic Drug Research 11
Shatrunjai P. Singh, Swagata Karkare, Sudhir M. Baswan, Vijendra P. Singh

Extended Marginal Homogeneity Model Based on Complementary Log-Log Transform for Square Tables 27
Yusuke Saigusa, Tomohisa Maruyama, Kouji Tahata, Sadao Tomizawa

Modeling Trend in Telecommunication in Sri Lanka: A Case study on Internet and Cellular Connections 32
K. A. N. K. Karunarathna, S. Brindha, P. Paramadevan

New Bounds on Poisson Approximation for Random Sums of Independent Binomial Random Variables 43
Giang Truong Le

On Optimal Allocation of Treatment/Condition Variance in Principal Component Analysis 50
André Beauducel, Norbert Hilger

The Two-Parameter Odd Lindley Weibull Lifetime Model with Properties and Applications 57
Jehhan. A. Almamy, Mohamed Ibrahim, M. S. Eliwa, Saeed Al-mualim, Haitham M. Yousof

The Impact of Sidewalks on Vehicle-Pedestrian Crash Severity 69
Mehrnaz Doustmohammadi, Niloufar Shirani Bidabadi, Sumalatha Kesaveraddy, Michael Anderson

Stress-Strength Reliability Model with The Exponentiated Weibull Distribution: Inferences and Applications 78
Fathy Helmy Eissa

The Bayes Factor for the Misclassified Categorical Data 91
Tze-San Lee

On Comparison of Local Polynomial Regression Estimators for P=0 and P=1 in a Model Based Framework 104
Conlet Biketi Kikechi, Richard Onyino Simwa

Reviewer Acknowledgements for International Journal of Statistics and Probability, Vol. 7, No. 4 115
Wendy Smith

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Explaining Lord's Paradox in Introductory Statistical Theory Courses

Steven B. Kim[1]

[1] Department of Mathematics and Statistics, California State University, Monterey Bay, USA

Correspondence: Steven B. Kim, Department of Mathematics and Statistics, California State University, Monterey Bay, Seaside, CA, USA.

Received: March 27, 2018 Accepted: April 14, 2018 Online Published: April 27, 2018

doi:10.5539/ijsp.v7n4p1 URL: https://doi.org/10.5539/ijsp.v7n4p1

Abstract

When two groups are compared in a pre-post study, two different conclusions can be drawn between the two-sample t-test and the analysis of covariance (ANCOVA). It is known as Lord's Paradox, and it occurs because the parameter in the two-sample t-test and the parameter of interest in the ANCOVA model are not the same quantity. The difference between the two parameters can be explained by the covariance of linearly combined random variables which is an important topic in introductory statistical theory courses. Lord's paradox is frequently observed in practice, and it is very important for students (future researchers) to have clear understanding of the paradox. The objective of this article is to explain Lord's Paradox using the covariance of linearly combined random variables. The paradox is explained using three scenarios in the context of educational research. The first scenario is when the average baseline (pre-score) is greater in the treatment group than the control group, the second scenario is when the average baseline is lower in the treatment group than the control group, and the third scenario is when the average baseline is same between the two groups by randomization. This article is written at the level of introductory statistical theory courses for undergraduate and graduate statistics students to help understanding the difference between the parameter of interest in the two-sample t-test and the parameter of interest in the ANCOVA model.

Keywords: two-sample t-test, ANCOVA, covariance, linear combination of random variables, pre-post studies

1. Introduction

When two groups are compared in a pre-post study, Lord's Paradox can be observed between two researchers when a researcher compares the average change using the two-sample t-test and the other researcher compares the average post-measurement using the analysis of covariance or simply ANCOVA (Lord 1967; Lord 1969). The paradox has been studied in the context of health sciences, environmental sciences, and psychometrics (Holland & Rubin, 1983; Wainer & Brown, 2006; Glymour et al., 2005; Tu et al., 2008; Pearl, 2016). It is an interesting phenomenon which frequently occurs in practice, but it is not easy to quantify the exact difference between the parameter in the two-sample t-test and the parameter in the ANCOVA model without statistical theory. In this article, we explain Lord's Paradox using the covariance of linearly combined random variables which is discussed in many statistical theory textbooks (Wackerly et al., 2008; Ross, 2012).

2. Motivating Example

The following example is adapted from the example given by Wright (2006). Suppose two groups of students are compared in their mathematics skills. Group 1 is the treatment group of size n_1 (receiving a new teaching method), and Group 0 is the control group of size n_0 (receiving a traditional teaching method). Assume each student took pre-test and post-test.

2.1 Scenario 1 (Wright, 2006)

Suppose each student selects a group by his or her own will. Suppose a student with high motivation (who tends to show high academic performance) is more likely to select Group 1, and suppose a student with relative low motivation is more likely to select Group 0. Wright (2006) illustrated a similar scenario with balanced group sizes $n_1 = 5$ and $n_0 = 5$ for Group 1 and Group 0, respectively. See Table 1 for the hypothetical data with minor modification from the example of Wright (2006).

Table 1. Hypothetical data of a pre-post study (Scenario 1)

ID	Group	Pre	Post	Difference
1	0	20	30	10
2	0	30	35	5
3	0	40	40	0
4	0	50	45	-5
5	0	60	50	-10
6	1	40	50	10
7	1	50	55	5
8	1	60	60	0
9	1	70	65	-5
10	1	80	70	-10

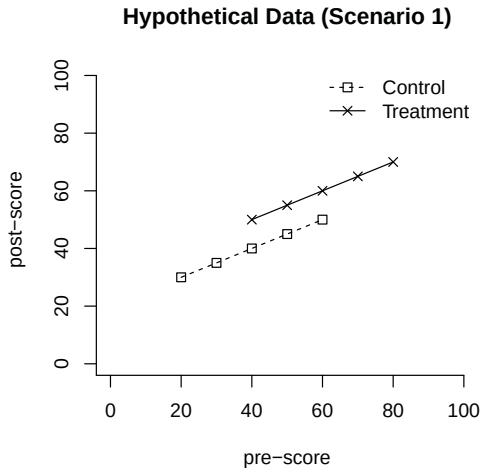


Figure 1. Hypothetical data of a pre-post study (Scenario 1)

The average difference is $(10 + 5 + 0 - 5 - 10) / 5 = 0$ for both groups which can be calculated from Table 1, but the post-score is 10 points greater on average when we condition on the pre-score as shown in Figure 1. (The data in real world may contain random noise around the line.) Using the two-sample t-test, the data are not against the null hypothesis at all (same group average). Using the ANCOVA model, on the other hand, the data are against the null hypothesis and serve as strong evidence for the alternative hypothesis (greater average post-score in Group 1 conditioning on pre-score). This is a traditional example of Lord's Paradox (Lord, 1967; Wright, 2003; Maxwell and Delaney, 2004; Wainer and Brown 2006). In addition to the graphic illustration, an analytic explanation of the paradox can be provided using the covariance of linearly combined random variables.

3. Covariance of Linearly Combined Random Variables

Several textbooks for the first semester of undergraduate statistical theory courses include the following proposition (Wackerly et al., 2008; Ross, 2012).

3.1 Proposition

Let $U_1, \ldots, U_n$ and $W_1, \ldots, W_m$ be random variables. Let $L_1 = \sum_{i=1}^{n} a_i U_i$ and $L_2 = \sum_{j=1}^{m} b_j W_j$ for fixed real numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$. Then

$$Cov(L_1, L_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \, Cov(U_i, W_j) \, .$$

Since $V(L_1) = Cov(L_1, L_1)$, a special result for the variance is

$$\begin{aligned} V(L_1) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \, Cov(U_i, U_j) \\ &= \sum_{i=1}^{n} a_i^2 V(U_i) + 2\sum_{i=1}^{n}\sum_{j>i} a_i a_j \, Cov(U_i, U_j)\,. \end{aligned}$$

From these results, we can explain why the two-sample t-test and the ANCOVA model can lead to different conclusions.

3.2 Two-sample t-test

Let Z_i denote the pre-score and Y_i denote the post-score of the i^{th} subject in a sample. Let X_i denote the group indicator for the i^{th} subject, where $X_i = 0$ for Group 0 (control) and $X_i = 1$ for Group 1 (treatment). The two-sample t-test can be formulated as a simple linear model

$$D_i = \beta_0 + \beta_1 X_i + \epsilon_i\,, \tag{1}$$

where $D_i = Y_i - Z_i$ is the change in test score (hence a positive value of D_i is a desirable outcome), and $\epsilon_i \sim N(0, \sigma^2)$ is a random variable which is independent of X_i. In Equation (1), the parameter of interest is the difference in the two group averages

$$\beta_1 = E(D_i \mid X_i = 1) - E(D_i \mid X_i = 0)\,.$$

The null hypothesis is H_0: $\beta_1 = 0$, and the one-sided alternative hypothesis is H_1: $\beta_1 > 0$. An alternative expression of β_1 is

$$\beta_1 = \frac{Cov(X_i, D_i)}{V(X_i)} \tag{2}$$

because

$$\begin{aligned} Cov(X_i, D_i) &= Cov(X_i, \beta_0 + \beta_1 X_i + \epsilon_i) \\ &= Cov(X_i, \beta_0) + Cov(X_i, \beta_1 X_i) + Cov(X_i, \epsilon_i) \\ &= \beta_1 \, V(X_i) \end{aligned}$$

by the proposition in Section 3.1.

3.3 ANCOVA

Preserving the same notation used in Section 3.2, the ANCOVA model assumes

$$Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \delta_i\,, \tag{3}$$

where $\delta_i \sim N(0, \tau^2)$ is a random variable which is independent of X_i and Z_i. Under the ANCOVA model, the parameter of interest is γ_1, the difference in the expected post-score when we compare a randomly selected subject in Group 1 to a randomly selected subject in Group 0 of the same pre-score. The null hypothesis is H_0: $\gamma_1 = 0$, and the one-sided alternative hypothesis is H_1: $\gamma_1 > 0$. An alternative expression of the ANCOVA model is

$$D_i = \gamma_0 + \gamma_1 X_i + (\gamma_2 - 1)\, Z_i + \delta_i$$

by subtracting Z_i on both sides of Equation (3). Using the proposition in Section 3.1,

$$\begin{aligned} Cov(X_i, D_i) &= Cov(X_i, \gamma_0 + \gamma_1 X_i + (\gamma_2 - 1)\, Z_i + \delta_i) \\ &= \gamma_1 V(X_i) + (\gamma_2 - 1)\, Cov(X_i, Z_i)\,, \end{aligned}$$

so the parameter of interest can be written as

$$\begin{aligned} \gamma_1 &= \frac{Cov(X_i, D_i) + (1 - \gamma_2)\, Cov(X_i, Z_i)}{V(X_i)} \\ &= \beta_1 + (1 - \gamma_2)\left(\frac{Cov(X_i, Z_i)}{V(X_i)}\right) \end{aligned}$$

from Equation (2). Using the same argument of the two-sample t-test, we can write

$$\kappa_1 \equiv \frac{Cov(X_i, Z_i)}{V(X_i)} = E(Z_i \mid X_i = 1) - E(Z_i \mid X_i = 0)\,,$$

which is interpreted as the difference in the average pre-score when we compare Group 1 to Group 0.

3.4 Summary

In general, the two-sample t-test and the ANCOVA model have different parameters of interest, and they are related as

$$\begin{aligned} \gamma_1 &= \beta_1 + (1-\gamma_2)\,\kappa_1\,, \\ \beta_1 &= \gamma_1 + (\gamma_2 - 1)\,\kappa_1\,. \end{aligned} \tag{4}$$

They are the same quantity (i.e., $\beta_1 = \gamma_1$) if $\kappa_1 = 0$ or $\gamma_2 = 1$. The first condition $\kappa_1 = 0$ can be satisfied by randomization (i.e., conducting an experimental study instead of an observational study), but the second condition $\gamma_2 = 1$ is out of researcher's control. In most pre-post studies, pre- and post-scores are positively correlated in both groups, so $\gamma_2 > 0$. In addition, we often have $0 < \gamma_2 < 1$ because of regression toward the mean (Stigler, 1997; Barnett et al., 2005).

4. Hypothetical Scenarios

In this section, using the relationship between β_1 and γ_1 in Equation (4), three scenarios are discussed in the context of the educational research. The first scenario is when the average baseline (pre-score) is greater in the treatment group than in the control group (Section 2.1), the second scenario is when the average baseline is lower in the treatment group than in the control group, and the third scenario is when the average baseline is the same between the treatment group and the control group by randomization. The control group is referred to as Group 0, and the treatment group is referred to as Group 1.

4.1 Revisiting Scenario 1

In Scenario 1 (from Section 2.1), the ordinary least square estimation (OLSE) results in $\hat{\gamma}_1 = 10$ and $\hat{\gamma}_2 = 0.5$. Due to self-selection by students, the pre-score is greater in Group 1 by 20 points on average when compared to Group 0, so

$$\hat{\beta}_1 = \hat{\gamma}_1 + (\hat{\gamma}_2 - 1)\,\hat{\kappa}_1 = 10 + (0.5 - 1)\,(20) = 0$$

for the two-sample t-test. This is an example of Lord's Paradox when the ANCOVA model can reject the null hypothesis, whereas the two-sample t-test cannot reject the null hypothesis even though the new teaching method seems significantly more effective than the traditional teaching method when we compare two randomly selected students from each group with the same baseline score.

4.2 Scenario 2 (Lower Average Baseline Score in the Treatment Group)

In the second scenario, assume the instructor allocates each student to Group 0 (control) or Group 1 (treatment) believing that the new teaching method would benefit students particularly with low academic performance. See Table 2 for hypothetical data, and see Figure 2 for the scatter plot of pre-score and post-score by group. Note that the pre-score is lower in Group 1 by 20 points on average when compared to Group 0 (i.e., $\hat{\kappa}_1 = -20$).

Table 2. Hypothetical data of a pre-post study (Scenario 2)

ID	Group	Pre	Post	Difference
1	0	40	45	5
2	0	50	50	0
3	0	60	55	-5
4	0	70	60	-10
5	0	80	65	-15
6	1	20	35	15
7	1	30	40	10
8	1	40	45	5
9	1	50	50	0
10	1	60	55	-5

From the data, the OLSE provides $\hat{\gamma}_1 = 0$ and $\hat{\gamma}_2 = 0.5$. In this scenario, the ANCOVA model cannot reject the null hypothesis because $\hat{\gamma}_1 = 0$. From Equation (4), for the two-sample t-test, we estimate $\hat{\beta}_1 = 0 + (0.5 - 1)\,(-20) = +10$ which can lead to the rejection of $\beta_1 = 0$ in favor of $\beta_1 > 0$ (i.e., greater benefit from the new teaching method). This is another example of Lord's Paradox when the two-sample t-test can reject the null hypothesis even though the new teaching method seems ineffective conditioning on the pre-score.

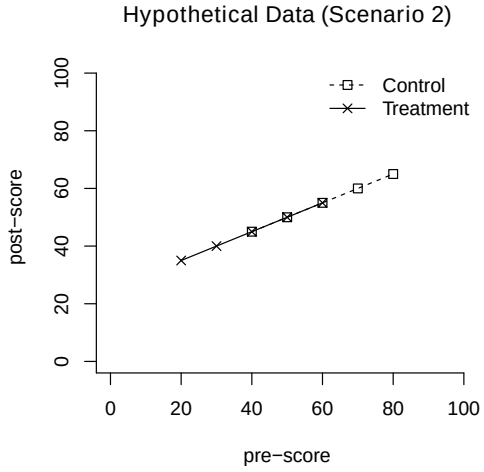


Figure 2. Hypothetical data of a pre-post study (Scenario 2)

4.3 Scenario 3 (Same Average Baseline Score between the Two Groups)

Suppose students are randomized (or controlled to match the average pre-score between the two groups) so that $\kappa_1 = 0$. In this case, the result from Equation (4) leads to $\beta_1 = \gamma_1$. As shown in Table 3 and Figure 3, we have $\hat{\kappa}_1 = 0$, so $\hat{\beta}_1 = \hat{\gamma}_1 = 10$, but the strength of statistical evidence for the alternative hypothesis is stronger in the ANCOVA model than in the two-sample t-test because the standard error is lower in the ANCOVA model. Though the ANCOVA model leads to nearly zero p-value, the two-sample t-test results in a p-value close to 0.05 (for the right-tail H_1: $\beta_1 > 0$). In practice, when students are randomized, the ANCOVA model should have higher statistical power than the two-sample t-test. It is because, while the OLSE is unbiased for both β_1 and γ_1, the variance of $Y_i - \gamma_2 Z_i$ is lower than the variance of $Y_i - Z_i$ conditioning on X_i as discussed in Appendix 1.

Table 3. Hypothetical data of a pre-post study (Scenario 3)

ID	Group	Pre	Post	Difference
1	0	30	40	10
2	0	40	45	5
3	0	50	50	0
4	0	60	55	-5
5	0	70	60	-10
6	1	30	50	20
7	1	40	55	15
8	1	50	60	10
9	1	60	65	5
10	1	70	70	0

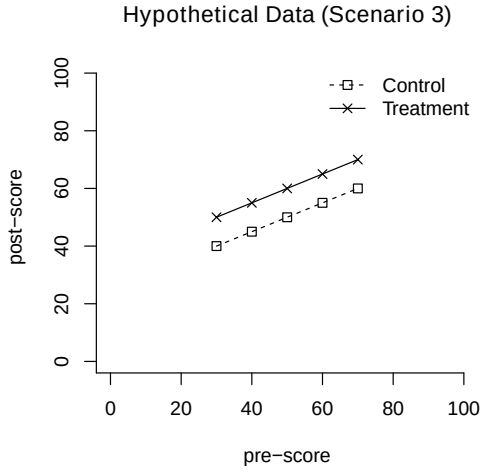


Figure 3. Hypothetical data of a pre-post study (Scenario 3)

5. Examples

In this section, we provide two practical examples. The example in Section 5.1 is to compare the effect of two programs on self-esteem score, and the example in Section 5.2 is to compare the effect of two teaching methods on test score.

5.1 Effect of Exercise on Self-Esteem

This example is from the data in R with `car` package (R Core Team, 2016; Fox & Weisberg, 2011). The data can be seen using the code below.

```
> library(car)
> WeightLoss
```

It has three groups, but we focus on two of the three groups. Twelve subjects ($n_0 = 12$) were treated by a diet program for three months, and this group is referred to as Group 0. Ten subjects ($n_1 = 10$) were treated by an exercise program in addition to the diet program for three months, and this group is referred to as Group 1. From the data presented in Table 4, we can estimate the average self-esteem score 14.8333 for Group 0 and 15.2 for Group 1 at Month 1, so $\hat{\kappa}_1 = 0.3667$.

To formulate hypothesis testing in terms of the expected change in self-esteem (comparing Month 3 to Month 1), the two-sample t-test can be used with H_0: $\beta_1 = 0$ versus H_1: $\beta_1 > 0$, assuming diet and exercise would be more beneficial than diet only, at significance level $\alpha = 0.05$. Using the two-sample t-test, we have a lack of evidence to reject H_0: $\beta_1 = 0$ with observed statistics $\hat{\beta}_1 = 1.0667$, $\widehat{\text{se}} = 0.6568$, $T = 1.624$, and p-value = 0.060.

To formulate hypothesis testing in terms of the expected self-esteem score at Month 3 given the score at Month 1, the ANCOVA model can be used with H_0: $\gamma_1 = 0$ versus H_1: $\gamma_1 > 0$ at $\alpha = 0.05$. Using the ANCOVA model, we have a statistically significance result to conclude H_1: $\gamma_1 > 0$ with observed statistics $\hat{\gamma}_1 = 1.1764$, $\widehat{\text{se}} = 0.6253$, $T = 1.881$, and p-value = 0.038.

In the left panel of Figure 4, the vertical distance between the two parallel lines is $\hat{\gamma}_1 = 1.1764$. In the right panel, the vertical distance between the two horizontal lines is $\hat{\beta}_1 = 1.0667$. Note that $\hat{\gamma}_2 = 0.7006$ in the ANCOVA model, and the estimated parameter in the two-sample t-test is slightly attenuated toward the null value $\beta_1 = 0$ because

$$\hat{\beta}_1 = \hat{\gamma}_1 + (\hat{\gamma}_2 - 1)\,\hat{\kappa}_1 = 1.1764 - (0.2994)(0.3667) = 1.0667$$

from Equation (4).

5.2 Comparing Two Teaching Methods

In a mathematics course, two teaching methods were compared for students' learning on set theory, and the learning was quantified by test scores. The first teaching method was based on a traditional lecture (Group 0), and the second teaching method was based on an active-based learning (Group 1). Each of twenty students was randomized into Group 0 or Group 1 by researchers ($n_0 = n_1 = 10$), and each student took a pre-test and a post-test on conceptual thinking.

The left panel of Figure 5 shows the pre-score on x-axis and the post-score on y-axis by group. Random numbers were

Table 4. Self-esteem data for comparing diet group (Group 0) and diet + exercise group (Group 1)

ID	Group (X_i)	Month 1 (Z_i)	Month 3 (Y_i)	Change (D_i)
1	0	12	14	+2
2	0	13	15	+2
3	0	17	18	+1
4	0	16	18	+2
5	0	16	15	–1
6	0	13	18	+5
7	0	12	14	+2
8	0	12	11	–1
9	0	17	19	+2
10	0	19	19	+0
11	0	15	15	+0
12	0	16	18	+2
13	1	15	19	+4
14	1	16	18	+2
15	1	13	17	+4
16	1	16	17	+1
17	1	13	16	+3
18	1	15	18	+3
19	1	15	18	+3
20	1	16	17	+1
21	1	16	19	+3
22	1	17	17	+0

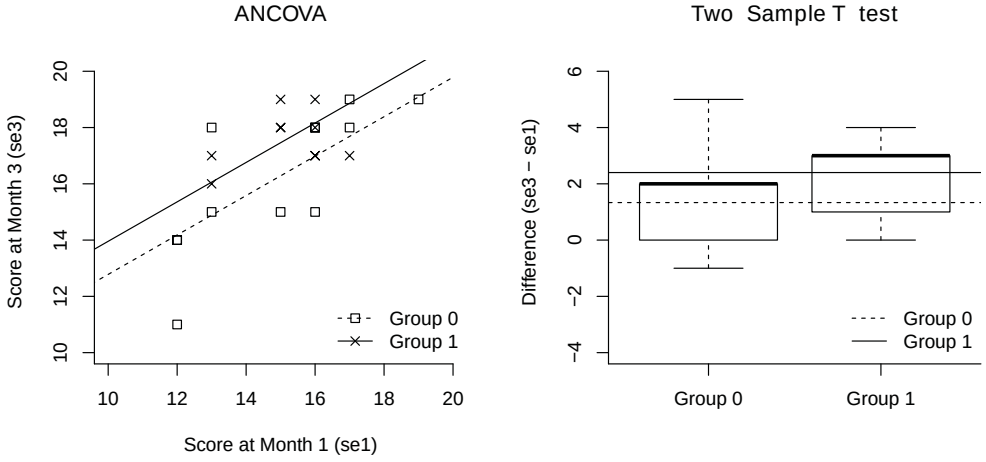


Figure 4. Data comparing diet group (Group 0) and diet + exercise group (Group 1)

generated by $N(0, \eta^2)$ with $\eta = 0.1$, and they were added to original data points for illustration purpose because it was difficult to show all twenty data points without the random noise. Under the ANCOVA model, we estimated $\hat{\gamma}_1 = 1.0283$ (with standard error $\widehat{\text{se}} = 0.3422$) and $\hat{\gamma}_2 = 0.2052$. For the hypothesis testing H_0: $\gamma_1 = 0$ and H_1: $\gamma_1 > 0$ at significance level $\alpha = 0.05$, we could reject H_0 in favor of H_1 with $T = 1.0283/0.3422 = 3.00$ and p-value 0.004.

The right panel of Figure 5 shows the difference in scores (post-score minus pre-score) by group, and the horizontal lines indicate the estimated average difference for each group. Despite the significant result from ANCOVA, the two boxplots look very similar except for one data point in Group 1. Even though the students were randomized, the difference in estimated average pre-score was $\hat{\kappa}_1 = 4.5209 - 3.8075 = 0.7134$ (comparing Group 1 to Group 0). From Equation (4), we can estimate $\hat{\beta}_1 = \hat{\gamma}_1 + (\hat{\gamma}_2 - 1)\,\hat{\kappa}_1 = 1.0283 - (0.7948)(0.7134) = 0.4613$. For the two-sample t-test, the estimated parameter $\hat{\beta}_1 = 0.4613$ was attenuated toward the null value $\beta_1 = 0$, the estimated standard error was $\widehat{\text{se}} = 0.5948$, and the resulting test statistic was $T = 0.4613/0.5948 = 0.776$ with p-value 0.224. Therefore, we could not reject H_0 in the

two-sample t-test at $\alpha = 0.05$.

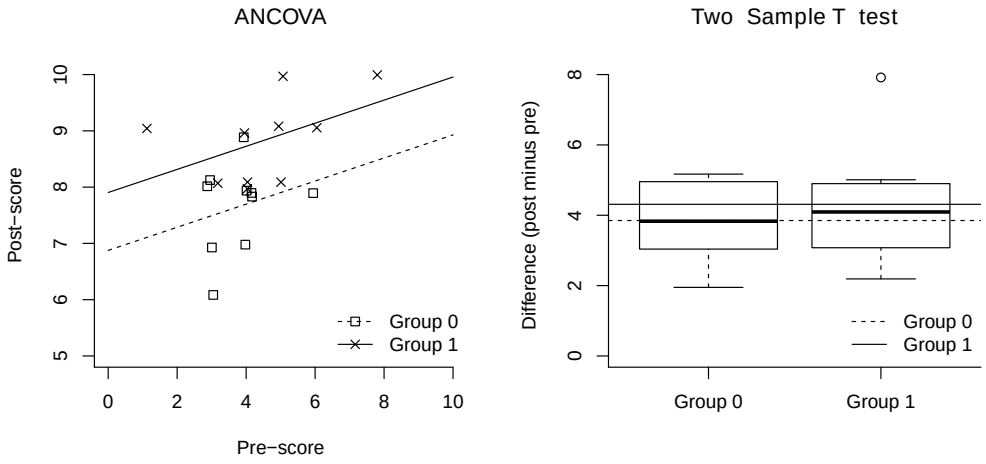


Figure 5. Data comparing traditional lecture (Group 0) and active-based learning (Group 1)

6. Discussion

Lord's Paradox has been known for a long time, and it has been explained graphically in literature, but it has received less attention analytically. Using the covariance of linearly combined random variables, we can show that the parameter β_1 in the two-sample t-test and the parameter γ_1 in the ANCOVA model are different by the magnitude of $(\gamma_2 - 1)\kappa_1$, where κ_1 is the difference in the average baseline score, comparing Group 1 (treatment) to Group 0 (control). In practice, it is difficult to have $(\gamma_2 - 1)\kappa_1 = 0$ in observational studies. This article can be summarized by the three scenarios in terms of the educational research scenarios presented in Section 4.

- When students with high baseline scores belong to the treatment group, which means $\kappa_1 > 0$, we have $\beta_1 < \gamma_1$. In an extreme case, we may have the opposite signs $\gamma_1 > 0$ and $\beta_1 < 0$.
- When students with low baseline scores belong to the treatment group, which means $\kappa_1 < 0$, we have $\beta_1 > \gamma_1$. When the treatment has no effect at all (i.e., H_0: $\gamma_1 = 0$ is true), there is a good chance of rejecting H_0: $\beta_1 = 0$ in favor of H_1: $\beta_1 > 0$ under the two-sample t-test with a large sample size.
- When students are randomized so that the average baseline score is same in the two groups, which means $\kappa_1 = 0$, we have $\beta_1 = \gamma_1$. In most practical situations, where pre- and post-scores are positively correlated in both groups, statistical power to conclude H_1: $\gamma_1 > 0$ in the ANCOVA model is greater than statistical power to conclude H_1: $\beta_1 > 0$ in the two-sample t-test as heuristically explained in Appendix 1.

The proposition in Section 3.1 is mentioned in most introductory statistical theory courses, and students can have deeper understanding of the two-sample t-test and the ANCOVA model through the examples.

In observational studies, we sometimes consider the propensity score, the conditional probability of assignment to a particular group (i.e., control or treatment) as a function of other variables, say $(W_1, \ldots, W_k)$ (Rosebaum & Rubin, 1983). The association between $(W_1, \ldots, W_k)$ and X_i does not necessarily imply the association between $(W_1, \ldots, W_k)$ and Y_i. In general, the difference between β_1 in the two-sample t-test and γ_1 in the multiple linear regression $Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \sum_{j=1}^{k} \alpha_j W_{j,i} + \delta_i$ can be quantified as $\beta_1 - \gamma_1 = (\gamma_2 - 1)\kappa_1 + \sum_{j=1}^{k} \alpha_j \nu_j$, where $\nu_j \equiv E(W_{j,i} \mid X_i = 1) - E(W_{j,i} \mid X_i = 0)$. See Appendix 2 for detail. If $W_{j,i}$ is not associated with Y_i given all other covariates (i.e., $\alpha_j = 0$), it does not contribute to the difference between β_1 and γ_1. The same argument holds for the use of a scalar propensity score, say S_i. The role of propensity score depends on the linear relationship between S_i and Y_i and $E(S_i \mid X_i = 1) - E(S_i \mid X_i = 0)$. Without any association between S_i and Y_i, the propensity score does not play any role in the difference between β_1 and γ_1.

References

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology, 34*, 215-220. https://doi.org/10.1093/ije/dyh299

Fox, J., & Weisberg, S. (2011). An R companion to applied regression (2nd edition). Thousand Oaks, CA: Sage.

Glymour, M. M., Weuve, J., Berkman, L. F., Kawachi, I., & Robins, J. M. (2005). When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *American Journal of Epidemilogy, 162*, 267-278. https://doi.org/10.1093/aje/kwi187

Holland, P., & Rubin, D. (1983). On Lord's Paradox. In Principals of Modern Psychological Measurement (H. Wainer and S. Messick, eds.). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304-305. http://dx.doi.org/10.1037/h0025105

Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin, 72*, 337-338. http://dx.doi.org/10.1037/h0028108

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analysing data: A model comparison perspective (2nd edition). Mahwah, NJ: Erlbaum.

Pearl, J. (2016). Lord's Paradox revisited C (Oh Lord! Kumbaya!). *Journal of Causal Inference, 4*(2). https://doi.org/10.1515/jci-2016-0021

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosebaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55. https://doi.org/10.1093/biomet/70.1.41

Ross, S. (2008). A first course in probability (8th edition). Upper Saddle River, NJ: Prentice Hall.

Stigler, S. (1997). Regression toward the mean, historically considered. *Statistical Methods in Medical Research, 6*, 103-114. https://doi.org/10.1177/096228029700600202

Tu, Y., Gunnell, D., & Gilthorpe, M. S. (2008). Simpson's paradox, Lord's Paradox, and suppression effects are the same phenomenon – the reversal paradox. *Emerging Themes in Epidemiology, 5*(2). https://doi.org/10.1186/1742-7622-5-2

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). Mathematical statistics with applications (7th edition). Belmont, London: Thomson Brooks/Cole.

Wainer, H., & Brown, L. M. (2006). Three statistical paradoxes in the interpretation of group differences: illustrated with medical school admission and licensing data. *Handbook of Statistics, 26*, 893-918. https://doi.org/10.1016/S0169-7161(06)26028-0

Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology, 73*, 123-136. https://doi.org/10.1348/000709903762869950

Wright, D. B. (2006). Comparing groups in a before-after design: when t test and ANCOVA produce different results. *British Journal of Educational Psychology, 76*, 663-675. https://doi.org/10.1348/000709905X52210

Appendix 1

In some sense, the two-sample t-test and the ANCOVA model have a common structure:

$$D_i = \beta_0 + \beta_1 X_i + \epsilon_i ,$$
$$D_i^* = \gamma_0 + \gamma_1 X_i + \delta_i ,$$

where $D_i = Y_i - Z_i$ in the two-sample t-test and $D_i^* = Y_i - \gamma_2 Z_i$ in the ANCOVA model. In hypothesis testing, when $\beta_1 = \gamma_1$, we can gain statistical power by having a smaller standard error (SE), and a lower SE can be achieved by a smaller variance of the dependent variable, D_i and D_i^*, given X_i. Assume subjects are randomized so that X_i and Z_i are

uncorrelated. Using the proposition in Section 3.1, we can express $V(D_i^*)$ as

$$\begin{aligned} V(D_i^*) &= V(Y_i - \gamma_2 Z_i) \\ &= V(Y_i) + \gamma_2^2\, V(Z_i) - 2\gamma_2\, Cov(Y_i, Z_i) \\ &= [V(Y_i) + V(Z_i) - 2\, Cov(Y_i, Z_i)] + \left[\gamma_2^2\, V(Z_i) - 2\gamma_2\, Cov(Y_i, Z_i) - V(Z_i) + 2\, Cov(Y_i, Z_i)\right] \\ &= V(D_i) - \left[(1-\gamma_2^2)\, V(Z_i) - 2(1-\gamma_2)\, Cov(Y_i, Z_i)\right], \end{aligned}$$

where

$$\begin{aligned} Cov(Y_i, Z_i) &= Cov(\gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \delta_i, Z_i) \\ &= \gamma_1\, Cov(X_i, Z_i) + \gamma_2\, V(Z_i) \\ &= \gamma_2\, V(Z_i) \end{aligned}$$

because $Cov(X_i, Z_i) = 0$ by the randomization. Therefore, we can simplify

$$\begin{aligned} V(D_i^*) &= V(D_i) - \left[(1-\gamma_2^2)\, V(Z_i) - 2(1-\gamma_2)\,\gamma_2\, V(Z_i)\right] \\ &= V(D_i) - (1-\gamma_2)\, V(Z_i)\, \left[(1+\gamma_2) - 2\gamma_2\right] \\ &= V(D_i) - (1-\gamma_2)^2\, V(Z_i)\,. \end{aligned}$$

To this end, we have $V(D_i^*) < V(D_i)$.

Appendix 2

In the two-sample T-test, the parameter of interest is

$$\beta_1 = \frac{Cov(X_i, D_i)}{V(X_i)} = E(D_i \mid X_i = 1) - E(D_i \mid X_i = 0)\,, \tag{5}$$

where $D_i = Y_i - Z_i$. If the multiple linear regression model is given by

$$Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \sum_{j=1}^{k} \alpha_j W_{j,i} + \delta_i\,,$$

we can write

$$D_i = \gamma_0 + \gamma_1 X_i + (\gamma_2 - 1)\, Z_i + \sum_{j=1}^{k} \alpha_j W_{j,i} + \delta_i\,.$$

Then the parameter of interest in the two-sample t-test is

$$\begin{aligned} \beta_1 &= \frac{\gamma_1 V(X_i) + (\gamma_2 - 1)\, Cov(X_i, Z_i) + \sum_{j=1}^{k} \alpha_j\, Cov(X_i, W_{j,i})}{V(X_i)} \\ &= \gamma_1 + (\gamma_2 - 1)\, \frac{Cov(X_i, Z_i)}{V(X_i)} + \sum_{j=1}^{k} \alpha_j\, \frac{Cov(X_i, W_{j,i})}{V(X_i)}\,. \end{aligned}$$

Since X_i is a Bernoulli random variable, as in Equation (5),

$$\begin{aligned} \kappa_1 &\equiv \frac{Cov(X_i, Z_i)}{V(X_i)} = E(Z_i \mid X_i = 1) - E(Z_i \mid X_i = 0)\,, \\ \nu_j &\equiv \frac{Cov(X_i, W_{j,i})}{V(X_i)} = E(W_{j,i} \mid X_i = 1) - E(W_{j,i} \mid X_i = 0) \end{aligned}$$

for $j = 1, \ldots, k$. Therefore,

$$\beta_1 = \gamma_1 + (\gamma_2 - 1)\,\kappa_1 + \sum_{j=1}^{k} \alpha_j \nu_j\,.$$

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The Application of Text Mining Algorithms In Summarizing Trends in Anti-Epileptic Drug Research

Shatrunjai P. Singh[1], Swagata Karkare[2], Sudhir M. Baswan[3] & Vijendra P. Singh[4]

[1] Lindner College of Business, University of Cincinnati, Ohio, USA

[2] School of Public Health, Boston University, Boston, MA, USA

[3] James L. Winkle College of Pharmacy, University of Cincinnati, Ohio, USA

[4] Department of Internal Medicine, Baptist Easley Hospital, Easley, SC, USA

Correspondence: Shatrunjai P. Singh, Lindner College of Business, University of Cincinnati, Ohio, USA.

Received: March 27, 2018 Accepted: April 19, 2018 Online Published: May 9, 2018

doi:10.5539/ijsp.v7n4p11 URL: https://doi.org/10.5539/ijsp.v7n4p11

Abstract

Content summarization is an important area of research in traditional data mining. The volume of studies published on anti-epileptic drugs (AED) has increased exponentially over the last two decades, making it an important area for the application of text mining based summarization algorithms. In the current study, we use text analytics algorithms to mine and summarize 10,000 PubMed abstracts related to anti-epileptic drugs published within the last 10 years. A Text Frequency – Inverse Document Frequency based filtering was applied to identify drugs with highest frequency of mentions within these abstracts. The US Food and Drug database was scrapped and linked to the results to quantify the most frequently mentioned modes of action and elucidate the pharmaceutical entities marketing these drugs. A sentiment analysis model was created to score the abstracts for sentiment positivity or negativity. Finally, a modified Latent Dirichlet Allocation topic model was generated to extract key topics associated with the most frequently mentioned AEDs. We found the top five most common drugs that appeared from the analysis were Gabapentin, Levetiracetam, Topiramate, Lamotrigine and Acetazolamide. We further listed the key topics associated with these drugs and the overall positive or negative sentiment associated with them. Results of this study provide accurate and data intensive insights on the progress of anti-epileptic drug research.

Keywords: Text Analytics, Anti-Epileptic Drugs, Sentiment Analysis, Topic Modeling

1. Introduction

The unparalleled surge in published biomedical literature has made it difficult to define quantitative and qualitative summarization of a specific topic. Recent advances in computational power have led to an increase in the use of text mining approaches to facilitate the summarization and content review (Khordad & Mercer, 2017; Moradi & Ghadiri, 2017; Zhu et al., 2013). Open source analytical tools can rapidly ingest vast sources and volumes of information which can then be further pipelined into key insights using algorithms like feature extraction, topic modeling and sentiment analysis, allowing accurate summarization (Mishra et al., 2014). These text mining approaches have already been employed in analyzing a wide array of topics like oncology databases (Zhu et al., 2013), impact of financial crises on suicides (Jung et al., 2017), awareness of climate change in rural communities (Bell, 2013) and analyzing the sentiment of diabetes patients on the twitter platform (Salas-Zárate et al., 2017). Although text mining has been employed in several domains of biomedical research, its use remains infrequent in many important therapeutic areas, including neuroscience research (Singh, 2015).

Epilepsy, the fourth most frequent neurological disorder, affects more than sixty million people globally (Singh, 2015; Singh & Karkare, 2017; Trinka et al., 2015; Singh, He, McNamara, & Danzer, 2013; Singh, LaSarge, An, McAuliffe, & Danzer, 2015). The social stigma linked with this condition often primes depression and is frequently associated with a decline in the quality of life (Benson et al., 2016; Hester et al., 2016; Luna et al., 2017). The problem is exacerbated by the confusion of focusing research efforts on multiple anti-epileptic drugs (AEDs), some of which show mixed results in the refractory epileptic populations (Ahmad et al., 2017; de Biase, Valente, Gigli, & Merlino, 2017; Nolan, Marson, Weston, & Tudur Smith, 2015; Pellock et al., 2017; Turner & Perry, 2017). The sheer volume of new research on AEDs cripples any meaningful insight generation.

In this study, we analyze 10,000 PubMed abstracts related to AEDs with the end goal of content summarization and

insight generation. Abstracts containing US FDA curated list of drugs were identified and analyzed for drug frequency, mode of action and the pharmaceutical entities manufacturing the most frequent drugs were acknowledged. A modified latent dirichlet analysis algorithm with a bigram tokenizer (mLDA) was used to extract key topics discussed in these abstracts. Finally, sentiment analysis was utilized to analyze which of these anti-epileptic drugs are promising candidates for further research based on associated positive sentiments.

2. Methods

2.1 Data Collection

We used an R-software based PubMed scrapper to download 10,000 abstracts positive either of these keywords: 'anti-epileptic drugs', 'anti-convulsant drugs' and 'AED'. Only abstracts published between 01/01/2007 to 01/01/2017 were included in the study. Papers with no abstracts or written in languages other than English were filtered out. The raw abstract data was uploaded to a public repository for open access (Singh & Karkare, 2018) . The raw R code used for the analysis was deposited in a *Github* repository (https://github.com/shatrunjai/aed_pubmed). A document term matrix (DTM) was created from the abstracts and was compared to the list of drugs approved in the last decade, obtained from the US Food and Drug Administration website (https://www.fda.gov/Drugs.htm). Only abstracts focusing on at least one of these drugs was included for further analysis.

2.2 Data Processing

Collected abstracts were scrubbed for numbers, non-English characters and stop words. The Stanford stop words list was used as the default stop word repository (https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms

-stop-words-1.html). Stemming of abstract was conducted according to the Porter stemmer (Porter, 1980). A document term matrix was created as described in Stanford NLP (https://nlp.stanford.edu/).A Term-Frequency-Inverse Document Frequency (TF-IDF) matrix was created and further frequency calculations were performed only on relevant TF-IDF terms as described in (Jones, 1972).The frequency matrix had a mean word frequency of 272 words and a standard deviation of 17 words. Words with frequency cut off two standard deviations from the mean word frequency were filtered from the list.

2.3 Modified LDA Based Topic Modelling

Latent Dirichlet Allocation (LDA) is a well-defined, unsupervised, generative, probabilistic method for modeling data and is frequently used in topic modeling (Blei, Ng, & Jordan, 2003). We created a modified Latent Dirichlet Allocation (mLDA) algorithm which assumes that each document can be denoted as a probabilistic distribution over latent topics and that the topic distribution in all documents share a common Dirichlet prior distribution. We also included a bigram tokenizer to better represent scientific abstracts. Each latent topic in the mLDA model is also represented as a probabilistic model over words and the word distributions of topics share a common Dirichlet prior distribution as well. Given a corpus M consisting of N documents, with document d having K_d words ($d \in \{1,..., N\}$), mLDA models M according to the following generative process (Blei et al., 2003; Li et al., 2016):

(a) Select a multinomial distribution φ_t for topic t ($t \in \{1,..., T\}$) from a Dirichlet prior distribution with parameter β.

(b) Select a multinomial distribution θ_d for document d ($d \in \{1,..., N\}$) from a Dirichlet prior distribution with parameter α.

(c) For a word w_n ($n \in \{1,..., K_d\}$) in document d,

(i) Select a topic z_n from θ_d.

(ii) Select a word w_n from φ_{zn}.

This generative process has words in documents are the only detected variables whereas others are latent variables (φ and θ) and hyper parameters (α and β). In order to deduce the latent variables and hyper parameters, the probability of experiential data M is calculated as follows:

$$p(M|\alpha,\beta) = \prod d = 1N \int p(\theta d|\alpha)\big(\sum n = 1Kdp(zdn|\theta d)p(wdn|zdn,\varphi)P(\varphi|\beta)\big)d\theta dd\varphi$$

Due to the coupling between θ and φ in the integrand (above equation), the precise implication in mLDA is obstinate (Blei et al., 2003).The number of topics was selected according to the Rate of Perplexity Change (RPC) previously described by Zhao and colleagues (Zhao et al., 2015). This algorithm yielded two key topics on average which were further curated manually.

2.4 Sentiment Analysis

To evaluate sentiment for each abstract, the *Sentiment Analysis* and *Tm* libraries were used within R-software ,Version 0.98.109 ("Text Mining Infrastructure in R | Feinerer | Journal of Statistical Software," n.d.). *Sentiment Analysis* algorithm is a well-established sentiment analysis (SA) protocol and has been cited by over a 1000 journal publications according to google scholar. *Sentiment Analysis* and *Tm* packages assign three sentiment scores ("positive," "negative," and "neutral") to each word, based on a generalized classification system developed by the authors which uses a combination of human-annotated and Artificial Intelligence based sentiment scoring algorithms (Bagheri & Islam, 2017). Further, we employed the "bag-of-words" approach which has been established to be very dependable for document-level SA, with aggregate-level performance approximately equivalent to more refined methods (Gayle & Shimaoka, 2017).

For the current study, nouns were excluded from the analysis as they contain little to no information (Pinheiro, Prado, Ferneda, & Ladeira, 2015). The sentiment of each abstract was calculated by combining the scores of all pertinent word tokens. A sentiment score ranging from −1 to +1 was allocated for each abstract based on the assessed grade of negative or positive sentiment. For further analysis and visualization, unstandardized scores were normalized to a distribution with a mean of zero ($\bar{x}=0$) and standard deviation of one ($\sigma_{\bar{x}}=1$). All abstracts were assigned values of 'positive' (score>+1),'negative' (score>-1) or 'neutral' (-1<score<+1).

2.5 Machine Learning

The *Sentiment Analysis* package uses a one class support vector machine (SVM) algorithm to classify the expressions and phrases within the abstracts based on Stanford core NLP trained algorithm. SVM is a supervised analytical method that classifies based on the degree to which the several input cases (i.e., expression vectors) predict a given binary class, like the presence of absence of positive sentiment (Salas-Zárate et al., 2017). All input terms, i.e. the bigrams can thus be assessed in terms of "importance" with respect to a given label (Gayle & Shimaoka, 2017).The classifier was retrained on a 7000-abstract sample curated dataset optimized for misclassification rate, precision and recall metrics.

2.6 Statistical Analyses

Microsoft SQL Server (version 2012) was used to query the dataset for different clone compositions, and statistical analysis was performed using R-statistical software (Version 0.98.109). Significance was determined using a two-tailed Student's t-test for data that met assumptions of normality and equal variance. The Mann-Whitney rank sum test was used for non-normal data. Proportions were compared using z-tests. Values presented are means ± SEM or medians [range], as appropriate. The experiment-wise error was conservatively set at 0.001 (Cumming, 2010). Corrections for multiple comparisons were done using a Bonferroni correction.

2.7 Figure Preparation

The results from R-software were exported into csv files which were imported into Tableau (version 8.0) or Microsoft Excel (version 2013) which were then used to create graphs and visualizations. Tables were created in Microsoft Word (version 2013).

3. Results

3.1 Characterizing the Most Published Anti-Epileptic Drugs in the Last 10 Years

To study AEDs that appeared in PubMed abstracts (2007-2017), an R scrapper was used to parse 10,000 PubMed abstracts. To identify abstracts specifically related to AEDs, this scrapped dataset was cross-referenced with the United States Food and National Drug database (US FDA) of drugs. A total of 130 drugs (Figure 1) with a mean of 69.34 abstracts per drugs and a standard deviation of 22.03 abstracts per drugs were identified. The top 5 most frequent drugs were: Gabapentin (abstract count=1371, Figure 1), Levetiracetam (abstract count=1304, Figure 1), Topiramate (abstract count=1027, Figure 1), Lamotrigine (abstract count=989, Figure 1) and Acetazolamide (abstract count=518, Figure 1). A year-by-year frequency of selected drug abstracts was performed for all the drugs beginning the year 1980 (Figure 2) to follow their research trends.

3.2 Characterizing Drug Class of the most Published Drugs

For all drugs, their pharmaceutical drug categorization was evaluated by using FDA definitions (https://www.fda.gov/drugs/informationondrugs/ucm079436.htm). As expected, Anti-Epileptic agents and CNS activity suppression agents were at the top of the list of our drug matches (Figure 3). However, cox-2 inhibitors, mood stabilizers, cytochrome p450-2C19 inhibitors, analgesics and serotonin reuptake inhibitors also frequent in the class of researched AEDs (Figures 3), reflecting the diversity in research initiatives.

3.3 Characterizing Pharmaceutical Industries with the Most Published Drugs

Next, the pharmaceutical companies associated with the highest frequency of drug mentions in the 10,000 abstracts selected for the study were extracted (Figure 4). Some companies had more than 5 drugs (Sagent Pharmaceuticals and Zydus Pharmaceuticals Inc. with 14 and 9 drugs, respectively). Zydus Pharmaceutical's Topiramate along with the other 8 drugs appears to lead the list in terms of the number of drugs and the frequency of abstract mentions. However, other companies like A-S Medication Solutions which despite having only one drug (Gabapentin), were still top-ranked in abstract mention frequency (Figure 4).

3.4 Using Sentiment Analysis to Score the Abstracts with the Top Anti-Epileptic Drugs

A sentiment analysis was performed on all the abstracts containing the keyword 'anti-epileptic drugs' or 'AED' or 'anti-convulsion drugs'. An initial analysis revealed a strong correlation between negative sentiment and the frequency of abstract mention (Table 1, correlation coefficient=0.68). To correct for this, a normalized sentiment score (Sentiment-Sentiment $_{mean}$/Sentiment $_{S.D.}$) was calculated for each drug (Table 1). The sentiment value/abstract correlation was manually tested for accuracy. Drugs Lisinopril (normalized sentiment score= -3.0) and Telmisartan (normalized sentiment score= 3.0) had the highest positive normalized sentiment of all the drugs, indicating that these appeared in abstracts with positive connotations ('positive outcome', 'no side effects') more often than other drugs. Conversely, Ethosuximide (normalized sentiment score= -0.9) and Meloxicam (normalized sentiment score= -2.3) had the most negative sentiment, indicating appearance in abstracts with negative connotations ('negative outcomes', 'side effects').

3.5 A modified Latent Dirichlet Algorithm Reveals Topics Associated with The Top 5 Most Mentioned Anti-Epileptic Drugs

An mLDA algorithm was employed to identify the key topics being discussed in the papers associated with the top 5 drug mentions (Table 2). Key words associated with the top topic indicated research on the lines of spinal surgery and pain outcomes. Levetiracetam was associated with topics including its use in refractory and generalized seizure, response bias by gender and its association with Brivaracetam. Topiramate was associated with topics including long term side effects, the development of drug-resistance, and its effect on Lennox-Gastaut syndrome. Acetazolamide was associated with one topic indicating research on its effect on visual acuity and macular degeneration. Finally, Lamotrigine was associated with one topic indicating possible side effects of dry mouth and blood spots at higher concentration of the drug.

4. Discussion

In this study, we use text analytics algorithms to summarize the latest development in anti-epileptic drug research. We mined the top five drugs that have been extensively published in PubMed, elucidate the pharmaceutical entities manufacturing/marketing these drugs, and also provided sentiment based direction on how this research is trending. Finally, we created an mLDA based topic modelling algorithm to discuss key topics associated with these drugs.

The most popular AED's conventionally used as first line treatment include primidone, ethosuximide, benzodiazepines, carbamazepine and phenobarbital. In the last 20 years, the Food and Drug Administration (FDA) has further approved twelve new AED's and have a longer list of these drugs in the clinical trial pipelines (Asconapé, 2010). Although all of these compounds have been used to treat epilepsy for more than a century a true anti-epileptic drug effective against all seizure grades and all demographics is still unavailable and approximately 30% of patients with epilepsy do not respond to any existing AEDs (Glauser et al., 2006; Singh, 2015). This has fueled basic research into new pharmacological agents with better safety and tolerability, ease of use and better titration rate, fewer potential interactions, and increased efficacy in comorbidities (Azar & Abou-Khalil, 2008). The resultant research from studies on different aspects of multiple AEDs has often made research summarization difficult and calls for newer computational approaches.

PubMed, the most extensively used warehouse of biomedical literature comprises of more than 20 million abstracts and is increasing at a frequency of over 90,000 abstracts per year: the quantity of articles added each year to PubMed has increased three times in the last 10 years (Andronis, Sharma, Virvilis, Deftereos, & Persidis, 2011). As research on a solitary subject may extend across numerous scientific areas and technical journals, it is progressively problematic for scientists to trail all advances in their area of work. The dispersal of information to many different journals and scientific subgroups has created and 'islets of scientific knowledge' and has led to the improvement of literature mining approaches pointing to link ideas and opinions that are not cited in the same editorial. The process of deducing implied knowledge from apparently unrelated concepts has been named literature-based discovery (LBD) (Andronis et al., 2011). These LBD methods have been used in the past for the purpose of theory ideation in association with drug discovery. Some of these LBD techniques include PubMed text mining, TF-IDF based keyword generation, unsupervised document clustering, literature modelling, sentiment analysis and topic modelling techniques. In the current study, we use a subset of these techniques for AED centered research summarization.

The most frequently studied AED was found to be Gabapentin, which is indicated for the treatment of postoperative

neuralgia in adults and for treating partial onset seizures in both pediatric and adult patients (Goa & Sorkin, 1993). Although the exact mode of Gabapentin action is unknown, it has been suggested that its activity depends on its interaction with voltage-gated calcium channels (Goa & Sorkin, 1993). Interestingly, topic modeling revealed the keywords 'pain' and 'spinal surgery' to be associated with this drug. However, although gabapentin is commonly used in pain management, its use in post-operative pain and spinal surgery is controversial (Chang, Challa, Shah, & Eloy, 2014; Singh, Singh, Fatima, Kubo, & Singh, 2008; Yu, Ran, Li, & Shi, 2013).

Levetiracetam, the second most commonly researched anti-epileptic drug, is indicated as an adjunctive therapy in the treatment of partial onset seizures in patients ≥16 years of age with epilepsy(Deshpande & Delorenzo, 2014; Zheng, Du, & Wang, 2015). The precise mechanism(s) by which Levetiracetam exerts its antiepileptic effect is unknown, but studies suggest that this agent acts as a neuromodulator and treats seizures by inhibiting presynaptic calcium channels (Deshpande & Delorenzo, 2014). Topic modeling from this study revealed recent efforts towards comparing the efficacy of Levetiracetam to Brivaricetam which has been a topic of increasing interest over the year (Crepeau & Treiman, 2010; Lyseng-Williamson, 2011).

Topiramate is used as a monotherapy in children of ages two and above and as an adjunctive therapy for adults. Its use is children is specifically indicated for seizures related with Lennox-Gastaut syndrome (LGS) (Crumrine, 2011; Donegan, Dixon, Hemming, Tudur-Smith, & Marson, 2015; Hoy, 2016). Topic modeling showed a strong association of this agent with Lennox-Gastaut syndrome, a disorder which initiates seizures in children (Crumrine, 2011; Singh, 2016; Singh et al., 2016; VanStraten & Ng, 2012).

Acetazolamide, a carbonic anhydrase inhibitor is indicated for the treatment of centrencephalic epilepsies (petit mal, unlocalized seizures) and is also a popular drug for the treatment of glaucoma (Reiss & Oles, 1996; Millichap & Aymat, 1967). Results of the topic modeling used in this study support a strong association of this drug with keywords like 'macular', 'visual', 'acuity', all of which are glaucoma-related terms referring to the discovery of its anti-epileptic properties during treatment of glaucoma patients (Lyall, 2008). Lamotrigine is an antiepileptic drug indicated as an adjunctive therapy in children above the ages of two specifically for primary generalized tonic-clonic seizures. It is also indicated for the treatment of bipolar disorder in patients (Ramaratnam, Panebianco, & Marson, 2016). Although the mechanism of action of this drug is unknown, in vitro pharmacological studies suggest that lamotrigine inhibits voltage-sensitive sodium channels, thereby stabilizing neuronal membranes and consequently modulating presynaptic transmitter release of excitatory amino acids (e.g., glutamate and aspartate). Topic modeling revealed the association of this drug with the terms 'dried blood spots', which suggests that research efforts have been focused on evaluating the safety profile of this drug, specifically in causing blood dyscrasias (Krasowski & McMillin, 2014; Milosheska, Grabnar, & Vovk, 2015; Baswan, Li, LaCount, & Kasting, 2016; Singh et al., 2016; Singh & Singh, 2017).

Sentiment analysis suggests that despite these drugs being well-established and approved lines of therapy in the treatment of a variety of epilepsies and seizures, all 5 drugs were associated with a negative sentiment. This indicates the possibility of mixed results in at least a subset of these research studies. Further, these findings suggest potential unmet need in the area of epilepsy treatment due to the dearth of positive sentiments surrounding these pharmacological agents.

5. Conclusion

This study demonstrates that although research efforts surrounding anti-epileptic treatments are moving in the right direction, there is an unmet need when it comes to the associated sentiments of researchers towards the most frequently studied agents. Despite the potential utility of these drugs in the treatment of epilepsy, their use in treatment could be hindered due to associated negative sentiments. Even though this study delineates the key topics surrounding AED research in the last decade, further research efforts should be conducted to understand the causal relationship between the negative sentiments and the pharmacological profile of these agents. Understanding these causative efforts can help lead the way for pharmaceutical manufacturers to devote research efforts towards improving the profiles of their drugs to better suit the needs of the patients.

References

Ahmad, K. A., Desai, S. J., Bennett, M. M., Ahmad, S. F., Ng, Y.-T., Clark, R. H., & Tolia, V. N. (2017). Changing antiepileptic drug use for seizures in US neonatal intensive care units from 2005 to 2014. *Journal of Perinatology: Official Journal of the California Perinatal Association*, *37*(3), 296–300. https://doi.org/10.1038/jp.2016.206

Andronis, C., Sharma, A., Virvilis, V., Deftereos, S., & Persidis, A. (2011). Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics*, *12*(4), 357–368. https://doi.org/10.1093/bib/bbr005

Asconapé, J. J. (2010). The selection of antiepileptic drugs for the treatment of epilepsy in children and adults. *Neurologic Clinics*, *28*(4), 843–852. https://doi.org/10.1016/j.ncl.2010.03.026

Azar, N. J., & Abou-Khalil, B. W. (2008). Considerations in the choice of an antiepileptic drug in the treatment of epilepsy.

Seminars in Neurology, *28*(3), 305–316. https://doi.org/10.1055/s-2008-1079335

Bagheri, H., & Islam, M. J. (2017). Sentiment analysis of twitter data. *ArXiv:1711.10377 [Cs]*. Retrieved from http://arxiv.org/abs/1711.10377

Baswan, S. M., Li, S. K., LaCount, T. D., & Kasting, G. B. (2016). Size and Charge Dependence of Ion Transport in Human Nail Plate. *Journal of Pharmaceutical Sciences*, *105*(3), 1201–1208. https://doi.org/10.1016/j.xphs.2015.12.011

Bell, E. J. (2013). Climate change and health research: has it served rural communities? *Rural and Remote Health*, *13*(1), 2343.

Benson, A., O'Toole, S., Lambert, V., Gallagher, P., Shahwan, A., & Austin, J. K. (2016). The stigma experiences and perceptions of families living with epilepsy: Implications for epilepsy-related communication within and external to the family unit. *Patient Education and Counseling*, *99*(9), 1473–1481. https://doi.org/10.1016/j.pec.2016.06.009

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Chang, C. Y., Challa, C. K., Shah, J., & Eloy, J. D. (2014). Gabapentin in acute postoperative pain management. *BioMed Research International*, *2014*, 631756. https://doi.org/10.1155/2014/631756

Crepeau, A. Z., & Treiman, D. M. (2010). Levetiracetam: a comprehensive review. *Expert Review of Neurotherapeutics*, *10*(2), 159–171. https://doi.org/10.1586/ern.10.5

Crumrine, P. K. (2011). Management of seizures in Lennox-Gastaut syndrome. *Paediatric Drugs*, *13*(2), 107–118. https://doi.org/10.2165/11536940-000000000-00000

Cumming, G. (2010). Replication, p rep, and confidence intervals: comment prompted by Iverson, Wagenmakers, and Lee (2010); Lecoutre, Lecoutre, and Poitevineau (2010); and Maraun and Gabriel (2010). *Psychological Methods*, *15*(2), 192–198. https://doi.org/10.1037/a0019521

de Biase, S., Valente, M., Gigli, G. L., & Merlino, G. (2017). Pharmacokinetic drug evaluation of lacosamide for the treatment of partial-onset seizures. *Expert Opinion on Drug Metabolism & Toxicology*, *13*(9), 997–1005. https://doi.org/10.1080/17425255.2017.1360278

Deshpande, L. S., & Delorenzo, R. J. (2014). Mechanisms of levetiracetam in the control of status epilepticus and epilepsy. *Frontiers in Neurology*, *5*, 11. https://doi.org/10.3389/fneur.2014.00011

Donegan, S., Dixon, P., Hemming, K., Tudur-Smith, C., & Marson, A. (2015). A systematic review of placebo-controlled trials of topiramate: How useful is a multiple-indications review for evaluating the adverse events of an antiepileptic drug? *Epilepsia*, *56*(12), 1910–1920. https://doi.org/10.1111/epi.13209

Gayle, A., & Shimaoka, M. (2017). Public Response to Scientific Misconduct: Assessing Changes in Public Sentiment Toward the Stimulus-Triggered Acquisition of Pluripotency (STAP) Cell Case via Twitter. *JMIR Public Health and Surveillance*, *3*(2). https://doi.org/10.2196/publichealth.5980

Glauser, T., Ben-Menachem, E., Bourgeois, B., Cnaan, A., Chadwick, D., Guerreiro, C., … Tomson, T. (2006). ILAE treatment guidelines: evidence-based analysis of antiepileptic drug efficacy and effectiveness as initial monotherapy for epileptic seizures and syndromes. *Epilepsia*, *47*(7), 1094–1120. https://doi.org/10.1111/j.1528-1167.2006.00585.x

Goa, K. L., & Sorkin, E. M. (1993). Gabapentin. A review of its pharmacological properties and clinical potential in epilepsy. *Drugs*, *46*(3), 409–427. https://doi.org/10.2165/00003495-199346030-00007

Hester, M. S., Hosford, B. E., Santos, V. R., Singh, S. P., Rolle, I. J., LaSarge, C. L., … Danzer, S. C. (2016). Impact of rapamycin on status epilepticus induced hippocampal pathology and weight gain. *Experimental Neurology*, *280*, 1–12. https://doi.org/10.1016/j.expneurol.2016.03.015

Hoy, S. M. (2016). Topiramate Extended Release: A Review in Epilepsy. *CNS Drugs*, *30*(6), 559–566. https://doi.org/10.1007/s40263-016-0344-5

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Khordad, M., & Mercer, R. E. (2017). Identifying genotype-phenotype relationships in biomedical text. *Journal of Biomedical Semantics*, *8*(1), 57. https://doi.org/10.1186/s13326-017-0163-8

Krasowski, M. D., & McMillin, G. A. (2014). Advances in anti-epileptic drug testing. *Clinica Chimica Acta; International Journal of Clinical Chemistry*, *436*, 224–236. https://doi.org/10.1016/j.cca.2014.06.002

Li, D., Wang, Z., Wang, L., Sohn, S., Shen, F., Murad, M. H., & Liu, H. (2016). A Text-Mining Framework for Supporting Systematic Reviews. *American Journal of Information Management*, *1*(1), 1–9.

Luna, J., Nizard, M., Becker, D., Gerard, D., Cruz, A., Ratsimbazafy, V., … Preux, P.-M. (2017). Epilepsy-associated levels of perceived stigma, their associations with treatment, and related factors: A cross-sectional study in urban and rural areas in Ecuador. *Epilepsy & Behavior: E&B*, *68*, 71–77. https://doi.org/10.1016/j.yebeh.2016.12.026

Lyall, D. A. M. (2008). Unexpected control of a patient's refractory epilepsy when treating glaucoma with acetazolamide. *Canadian Journal of Ophthalmology. Journal Canadien D'ophtalmologie*, *43*(3), 377. https://doi.org/10.3129/i08-036

Lyseng-Williamson, K. A. (2011). Levetiracetam: a review of its use in epilepsy. *Drugs*, *71*(4), 489–514. https://doi.org/10.2165/11204490-000000000-00000

Millichap, J. G., & Aymat, F. (1967). Treatment and prognosis of petit mal epilepsy. *Pediatric Clinics of North America*, *14*(4), 905–920.

Milosheska, D., Grabnar, I., & Vovk, T. (2015). Dried blood spots for monitoring and individualization of antiepileptic drug treatment. *European Journal of Pharmaceutical Sciences: Official Journal of the European Federation for Pharmaceutical Sciences*, *75*, 25–39. https://doi.org/10.1016/j.ejps.2015.04.008

Mishra, R., Bian, J., Fiszman, M., Weir, C. R., Jonnalagadda, S., Mostafa, J., & Del Fiol, G. (2014). Text summarization in the biomedical domain: a systematic review of recent research. *Journal of Biomedical Informatics*, *52*, 457–467. https://doi.org/10.1016/j.jbi.2014.06.009

Moradi, M., & Ghadiri, N. (2017). Different approaches for identifying important concepts in probabilistic biomedical text summarization. *Artificial Intelligence in Medicine*. https://doi.org/10.1016/j.artmed.2017.11.004

Nolan, S. J., Marson, A. G., Weston, J., & Tudur Smith, C. (2015). Carbamazepine versus phenobarbitone monotherapy for epilepsy: an individual participant data review. *The Cochrane Database of Systematic Reviews*, (7), CD001904. https://doi.org/10.1002/14651858.CD001904.pub2

Pellock, J. M., Arzimanoglou, A., D'Cruz, O., Holmes, G. L., Nordli, D., Shinnar, S., & Pediatric Epilepsy Academic Consortium for Extrapolation. (2017). Extrapolating evidence of antiepileptic drug efficacy in adults to children ≥2 years of age with focal seizures: The case for disease similarity. *Epilepsia*, *58*(10), 1686–1696. https://doi.org/10.1111/epi.13859

Pinheiro, M. S., Prado, H. A. do, Ferneda, E., & Ladeira, M. (2015). An Approach for Text Mining Based on Noun Phrases. In *Intelligent Decision Technologies* (pp. 525–535). Springer, Cham. https://doi.org/10.1007/978-3-319-19857-6_45

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Ramaratnam, S., Panebianco, M., & Marson, A. G. (2016). Lamotrigine add-on for drug-resistant partial epilepsy. *The Cochrane Database of Systematic Reviews*, (6), CD001909. https://doi.org/10.1002/14651858.CD001909.pub2

Reiss, W. G., & Oles, K. S. (1996). Acetazolamide in the treatment of seizures. *The Annals of Pharmacotherapy*, *30*(5), 514–519. https://doi.org/10.1177/106002809603000515

Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á., & Valencia-García, R. (2017). Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. *Computational and Mathematical Methods in Medicine*, *2017*, 5140631. https://doi.org/10.1155/2017/5140631

Singh, S. P. (2015). *Quantitative analysis on the origins of morphologically abnormal cells in temporal lobe epilepsy*. University of Cincinnati. Retrieved from https://etd.ohiolink.edu/pg_10?0::NO:10:P10_ACCESSION_NUM:ucin1446547280

Singh, S. P. (2016). Advances in Epilepsy: A data science perspective, *58*(2), 89–92. https://doi.org/10.2791/dsj.7.1

Singh, S. P., Chhunchha, B., Fatma, N., Kubo, E., Singh, S. P., & Singh, D. P. (2016). Delivery of a protein transduction domain-mediated Prdx6 protein ameliorates oxidative stress-induced injury in human and mouse neuronal cells. *American Journal of Physiology. Cell Physiology*, *310*(1), C1-16. https://doi.org/10.1152/ajpcell.00229.2015

Singh, S. P., He, X., McNamara, J. O., & Danzer, S. C. (2013). Morphological changes among hippocampal dentate granule cells exposed to early kindling-epileptogenesis. *Hippocampus*, *23*(12), 1309–1320. https://doi.org/10.1002/hipo.22169

Singh, S. P., & Karkare, S. (2017). Stress, Depression and Neuroplasticity. *ArXiv:1711.09536 [q-Bio]*. Retrieved from http://arxiv.org/abs/1711.09536

Singh, S. P., & Karkare, S. (2018, January 5). 10K Pubmed Abstracts related to AntiEpileptic Drugs. https://doi.org/10.6084/m9.figshare.5764524.v1

Singh, S. P., LaSarge, C. L., An, A., McAuliffe, J. J., & Danzer, S. C. (2015). Clonal Analysis of Newborn Hippocampal Dentate Granule Cell Proliferation and Development in Temporal Lobe Epilepsy. *Eneuro*, *2*(6). https://doi.org/10.1523/ENEURO.0087-15.2015

Singh, S. P., Singh, S. P., Fatima, N., Kubo, E., & Singh, D. P. (2008). Peroxiredoxin 6-A novel antioxidant neuroprotective agent. *Neurology*, *70*(11), A480–A481.

Singh, S. P., & Singh, V. P. (2017). Quantitative Analysis on the role of Raffinose Synthase in Hippocampal Neurons. *BioRxiv*. https://doi.org/10.1101/240192

Text Mining Infrastructure in R | Feinerer | Journal of Statistical Software. (n.d.). https://doi.org/10.18637/jss.v025.i05

Trinka, E., Cock, H., Hesdorffer, D., Rossetti, A. O., Scheffer, I. E., Shinnar, S., … Lowenstein, D. H. (2015). A definition and classification of status epilepticus--Report of the ILAE Task Force on Classification of Status Epilepticus. *Epilepsia*, *56*(10), 1515–1523. https://doi.org/10.1111/epi.13121

Turner, A. L., & Perry, M. S. (2017). Outside the box: Medications worth considering when traditional antiepileptic drugs have failed. *Seizure*, *50*, 173–185. https://doi.org/10.1016/j.seizure.2017.06.022

VanStraten, A. F., & Ng, Y.-T. (2012). Update on the management of Lennox-Gastaut syndrome. *Pediatric Neurology*, *47*(3), 153–161. https://doi.org/10.1016/j.pediatrneurol.2012.05.001

Yu, L., Ran, B., Li, M., & Shi, Z. (2013). Gabapentin and pregabalin in the management of postoperative pain after lumbar spinal surgery: a systematic review and meta-analysis. *Spine*, *38*(22), 1947–1952. https://doi.org/10.1097/BRS.0b013e3182a69b90

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, *16*(Suppl 13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8

Zheng, F., Du, C., & Wang, X. (2015). Levetiracetam for the treatment of status epilepticus. *Expert Review of Neurotherapeutics*, *15*(10), 1113–1121. https://doi.org/10.1586/14737175.2015.1088785

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., … Shen, B. (2013). Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*, *46*(2), 200–211. https://doi.org/10.1016/j.jbi.2012.10.007

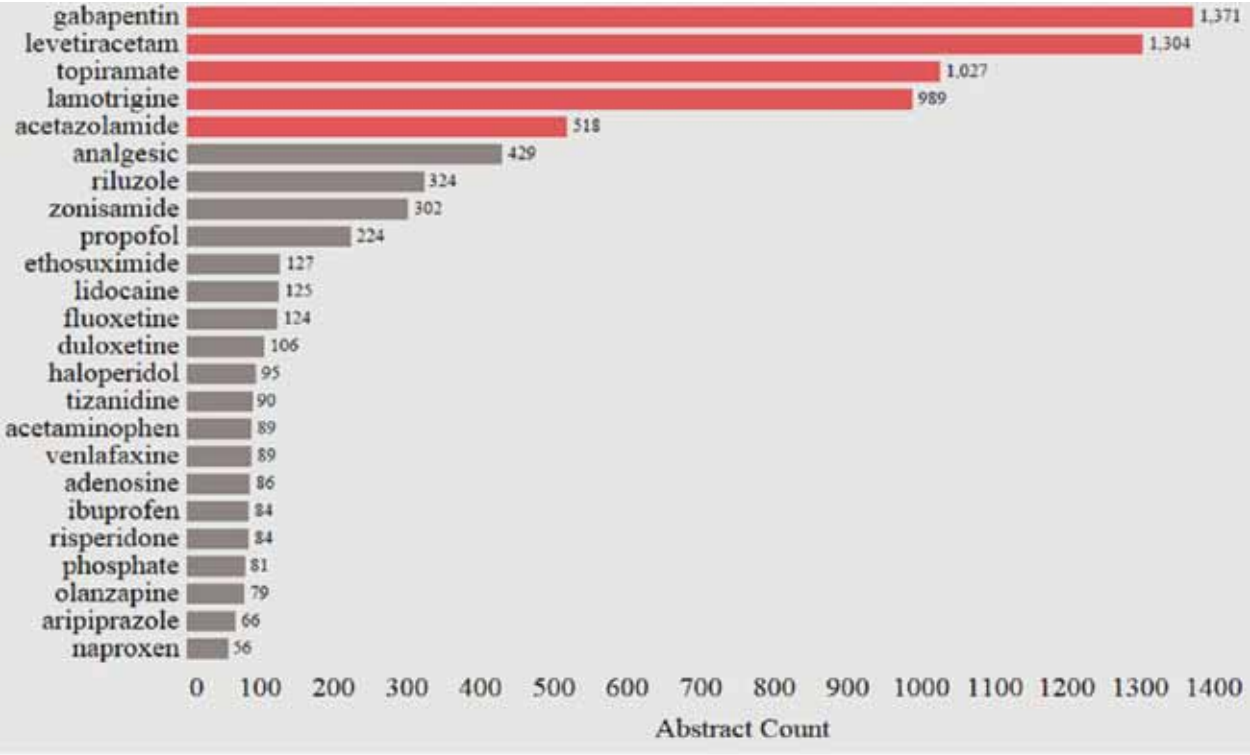


Figure 1: The frequency of PubMed abstracts with anti-epileptic drug mentions from 2007 to 2017. Shown here are the drugs with more than 50 abstracts.

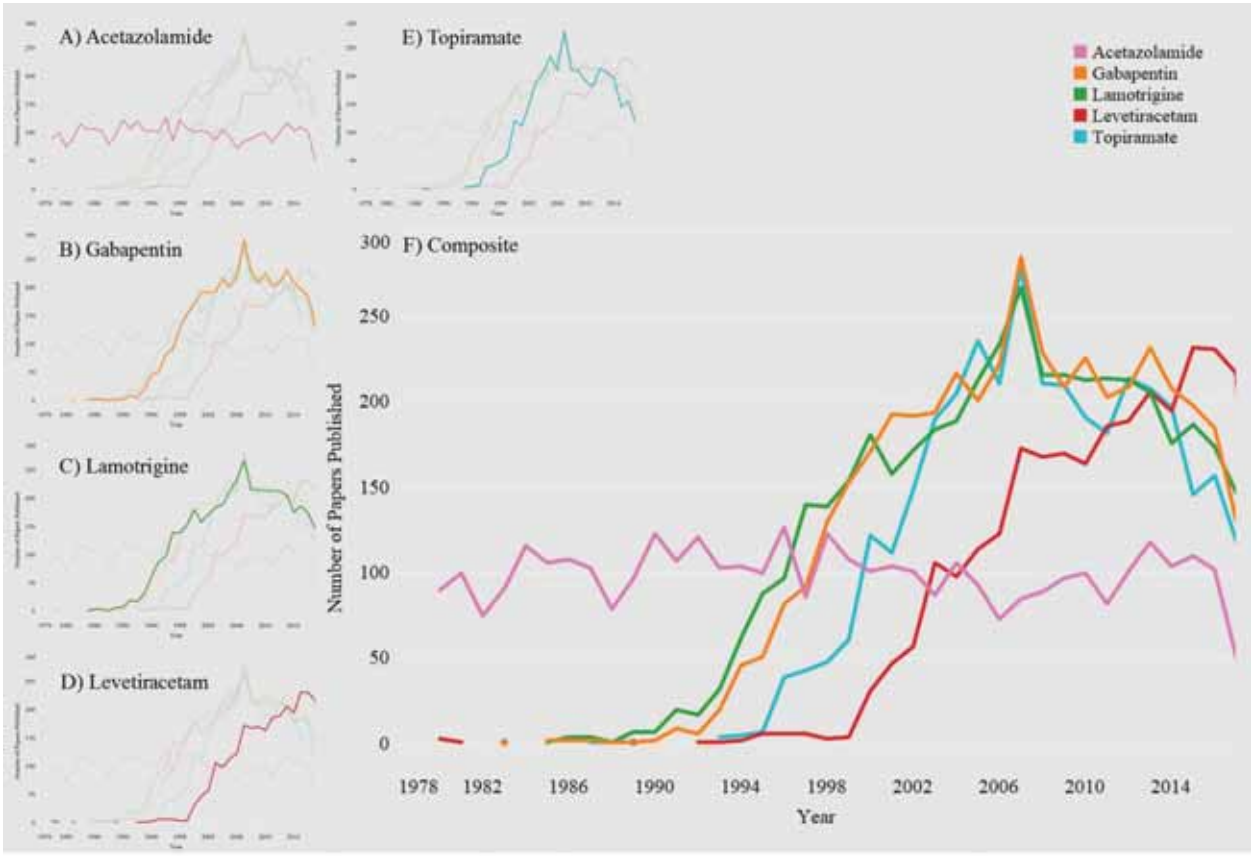


Figure 2: Yearly mentions of the five most frequently occurring anti-epileptic drugs. The frequency of abstracts published between 1980 to 2017 for A) Acetazolamide, B) Gabapentin, C) Lamotrigine, D) Levetiracetam and E) Topiramate and F) Composite of all the drugs (A through E)

Figure 3 The most common drug classes found in the abstract with the frequency of their occurrence. Drug classification was obtained from the US FDA classification.

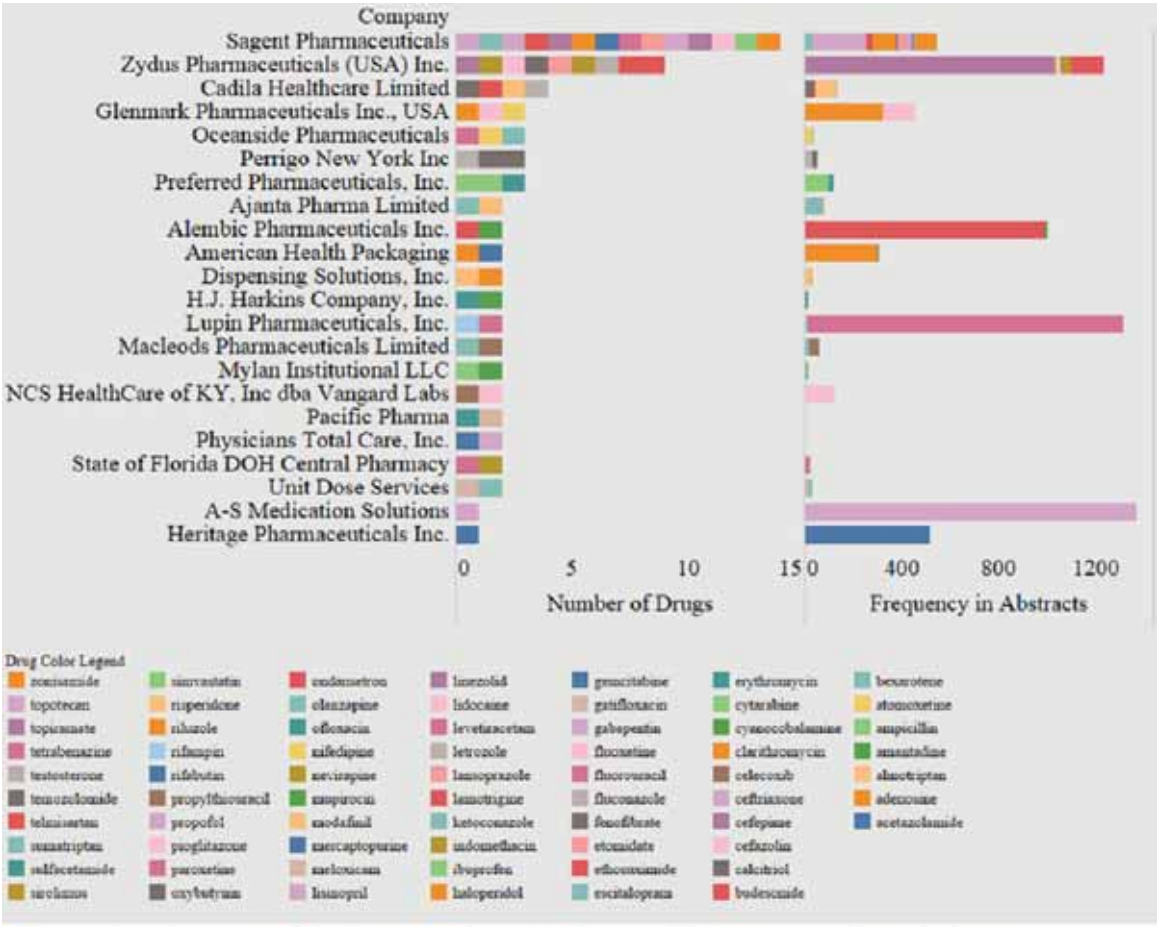


Figure 4. Pharmaceutical companies which have the most number of drugs mentioned in the last 10000 abstracts. The registered name of the company is shown in the first pane, the number of drugs that made it into the study criterion in the second pane and the frequency of drug appearance in the abstracts in the third pane.

Table 1. Sentiment analysis on abstracts with AED mentions. The sentiment of all the abstracts (not-normalized) is shown as either positive (score>0) or negative (score<0). The normalized sentiment scores (score-mean/S.D.) is shown in the right most column. The mean sentiment score was 0.01320 and the standard deviation was 0.2772.

Drug	Sentiment Polarity	Sentiment	Normalized Sentiment
Clopidogrel	Positive	0.002897016	-0.037168052
Mirtazapine	Positive	0.005795628	-0.026711299
Topotecan	Positive	0.009118557	-0.01472382
Fenofibrate	Positive	0.01211505	-0.003913961
Methazolamide	Positive	0.01266136	-0.001943146
Gemcitabine	Positive	0.01526285	0.007441739
Cimetidine	Positive	0.01862922	0.019585931
Cilostazol	Positive	0.02063282	0.026813925
Leflunomide	Positive	0.02083333	0.027537266
Fluoxetine	Positive	0.02222189	0.032546501
Epinephrine	Positive	0.02343064	0.036907071
Loratadine	Positive	0.02425356	0.039875758
Sulfacetamide	Positive	0.02611344	0.046585281
Antibacterial	Positive	0.03498982	0.078606854
Fluorouracil	Positive	0.039489	0.094837662
Ribavirin	Positive	0.03993905	0.096461219
Testosterone	Positive	0.04173582	0.102943074
Gatifloxacin	Positive	0.04434233	0.112346068
Decitabine	Positive	0.04531584	0.115858009
Rifabutin	Positive	0.04868627	0.128016847
Tetrabenazine	Positive	0.05354119	0.145530988
Pioglitazone	Positive	0.05555496	0.152795671
Carisoprodol	Positive	0.05842708	0.163156854
Cytarabine	Positive	0.05848246	0.163356638
Misoprostol	Positive	0.0597341	0.167871934
Omeprazole	Positive	0.06585542	0.189954618
Budesonide	Positive	0.06689463	0.193703571
Calcitriol	Positive	0.06788645	0.197281566
Ketoconazole	Positive	0.08378791	0.25464614
Venlafaxine	Positive	0.0912101	0.281421717
Ganciclovir	Positive	0.09387124	0.291021789
Clarithromycin	Positive	0.09424993	0.292387915
Sumatriptan	Positive	0.1025025	0.322159091
Rifampin	Positive	0.1035337	0.325879149
Piroxicam	Positive	0.1071953	0.339088384
Temozolomide	Positive	0.1075685	0.340434704
Almotriptan	Positive	0.1216465	0.39122114

Itraconazole	Positive	0.1348309	0.438783911
Linezolid	Positive	0.1455073	0.477299062
Sirolimus	Positive	0.148724	0.488903319
Voriconazole	Positive	0.1552595	0.512480159
Paroxetine	Positive	0.1920355	0.645149711
Oxybutynin	Positive	0.20757	0.701190476
Escitalopram	Positive	0.2461041	0.840202381
Propylthiouracil	Positive	0.3178791	1.099130952
Oxandrolone	Positive	0.3825442	1.332410534
Mercaptopurine	Positive	0.4220309	1.474858947
Chlorzoxazone	Positive	0.5641749	1.987643939
Telmisartan	Positive	1.026542	3.655634921
Lisinopril	Positive	2.534934	9.097164502
Meloxicam	Negative	-0.6005493	-2.214102814
Ethosuximide	Negative	-0.2520805	-0.957000361
Hemorrhoidal	Negative	-0.2358386	-0.898407648
Lansoprazole	Negative	-0.2164586	-0.828494228
Cefazolin	Negative	-0.2010915	-0.773057359
Laxative	Negative	-0.2001804	-0.769770563
Expectorant	Negative	-0.1859978	-0.718606782
Menthol	Negative	-0.1659774	-0.646383117
Cefepime	Negative	-0.1627273	-0.634658369
Celecoxib	Negative	-0.1516571	-0.594722583
Olanzapine	Negative	-0.1500488	-0.588920635
Bacitracin	Negative	-0.1498865	-0.588335137
Atomoxetine	Negative	-0.1445815	-0.56919733
Bexarotene	Negative	-0.1382776	-0.546455988
Bupropion	Negative	-0.1250089	-0.498589105
Lidocaine	Negative	-0.1211533	-0.484680014
Cyanocobalamine	Negative	-0.1130527	-0.455457071
Ampicillin	Negative	-0.1123144	-0.452793651
Indomethacin	Negative	-0.1115034	-0.449867965
Carboplatin	Negative	-0.1075859	-0.43573557
Gabapentin	Negative	-0.107477	-0.435342713
Amantadine	Negative	-0.105775	-0.429202742
Oxaliplatin	Negative	-0.1030068	-0.41921645
Aspirin	Negative	-0.1023196	-0.416737374
Ceftriaxone	Negative	-0.1000455	-0.40853355
Diphenhydramine	Negative	-0.09806558	-0.401390981
Cortisone	Negative	-0.09325419	-0.384033874
Modafinil	Negative	-0.09214975	-0.380049603
Doxycycline	Negative	-0.08559167	-0.356391306

Clotrimazole	Negative	-0.0839393	-0.350430375
Metronidazole	Negative	-0.08369748	-0.349558009
Letrozole	Negative	-0.08360991	-0.3492421
Paclitaxel	Negative	-0.08267615	-0.345873557
Topiramate	Negative	-0.08196571	-0.343310642
Lamotrigine	Negative	-0.07665585	-0.324155303
Temazepam	Negative	-0.07443998	-0.316161544
Azithromycin	Negative	-0.07425097	-0.31547969
Hydrocortisone	Negative	-0.07420852	-0.315326551
Valsartan	Negative	-0.0687518	-0.295641414
Ciprofloxacin	Negative	-0.06780962	-0.292242496
Adenosine	Negative	-0.06476878	-0.281272655
Risperidone	Negative	-0.06339783	-0.276326948
Etomidate	Negative	-0.05926828	-0.261429582
Aripiprazole	Negative	-0.05923076	-0.261294228
Nicotine	Negative	-0.05807075	-0.257109488
Zaleplon	Negative	-0.0561836	-0.250301587
Duloxetine	Negative	-0.05339635	-0.240246573
Tizanidine	Negative	-0.05224761	-0.236102489
Acetaminophen	Negative	-0.05086406	-0.231111328
Ropinirole	Negative	-0.0489206	-0.224100289
Hydrochlorothiazide	Negative	-0.04238263	-0.200514538
Levetiracetam	Negative	-0.04194175	-0.198924062
Phosphate	Negative	-0.04153744	-0.197465512
Nevirapine	Negative	-0.04135263	-0.19679881
Guanfacine	Negative	-0.03962207	-0.190555808
Isoniazid	Negative	-0.0354021	-0.175332251
Simvastatin	Negative	-0.03430078	-0.171359235
Acetazolamide	Negative	-0.0340079	-0.17030267
Propofol	Negative	-0.03290462	-0.166322583
Ondansetron	Negative	-0.03218649	-0.163731926
Ofloxacin	Negative	-0.03167064	-0.161870996
Levofloxacin	Negative	-0.02956457	-0.154273341
Zonisamide	Negative	-0.02831881	-0.149779257
Riluzole	Negative	-0.02446827	-0.13588842
Nifedipine	Negative	-0.02092019	-0.123088709
Furosemide	Negative	-0.01862298	-0.114801515
Ibuprofen	Negative	-0.01638219	-0.106717857
Naproxen	Negative	-0.01059067	-0.085824928
Fluconazole	Negative	-0.007489149	-0.07463618
Eszopiclone	Negative	-0.006686009	-0.071738849
Haloperidol	Negative	-0.005274025	-0.066645112

Disposable	Negative	-0.004070103	-0.062301959
Minoxidil	Negative	-0.001773493	-0.05401693
Lovastatin	Negative	-0.001518158	-0.053095808
Acyclovir	Negative	-0.001015911	-0.05128395
Erythromycin	Negative	-0.00086424	-0.050736798

Table 2. Results from mLDA based Topic models run on abstract containing the top 5 drugs. Each topic is represented by the top keywords defining the topic.

Gabapentin		**Levetiracetam**		**Topiramate**		**Acetazolamide**	**Lamotrigine**
Topic 1	**Topic 2**	**Topic 1**	**Topic 2**	**Topic 1**	**Topic 2**	**Topic 1**	**Topic 1**
Pain	Analog	Refractory	Indicated	LGS	Controlling	Clinical	Blood
Group	Incidence	Generalized	Woman	Epilepsy	Drug Resistant	Macular	Dried
Controlled	Scores	Seizures	Partial Epilepsy	Lennoxgastaut	Effective	Visual	Spots
Spinal	Inclusion	New	Response	Drugs	Long Term	Acuity	Concentrations
Surgery	Scale	Brivaracetam	Therapy	Resolution	Drop	Better	Accuracy

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Extended Marginal Homogeneity Model Based on Complementary Log-Log Transform for Square Tables

Yusuke Saigusa[1], Tomohisa Maruyama[2], Kouji Tahata[2] & Sadao Tomizawa[2]

[1] Department of Biostatistics, Yokohama City University School of Medicine, Japan

[2] Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Japan

Correspondence: Yusuke Saigusa, Department of Biostatistics, Yokohama City University School of Medicine, Yokohama, Kanagawa 236–0004, Japan.

Received: April 2, 2018 Accepted: April 17, 2018 Online Published: May 28, 2018

doi:10.5539/ijsp.v7n4p27 URL: https://doi.org/10.5539/ijsp.v7n4p27

Abstract

For square contingency tables with the same ordinal row and column classifications, McCullagh (1977) gave the marginal cumulative logistic model, which is an extension of the marginal homogeneity (MH) model using the logit transform. The present paper proposes a different extension of the MH model using the complementary log-log transform. In addition, the present paper gives the theorem that the MH model is equivalent to the proposed model and the equality of row and column marginal means holding simultaneously. In data analysis, if the MH model fits the data poorly, the theorem may be useful for seeing the reason for the poor fit. As example, the occupational status data for British father-son pairs are analyzed.

Keywords: decomposition, mean equality, logit transform

1. Introduction

Consider a square contingency table with the same ordinal row and column classifications. In the data in Table 1 taken from Bishop, Fienberg & Holland (1975, p.100), each observation is a pair of father's occupational status with his son's occupational status. For such data, statistical independence between the row and column classification generally does not hold due to concentration of observations on main diagonal cells. Instead of independence, we are interested in whether there is a structure of symmetry in the table. For example, Stuart (1955) gave the marginal homogeneity (MH) model which states the row marginal distribution is identical to the column marginal distribution. It is known that the MH model is expressed as the equality of marginal cumulative probabilities of row and column. For the data in Table 1, the MH model indicates the probability that a father's status is i equals the probability that his son's status is also i for any category i.

In data analysis, when the MH model fits the data poorly, many statisticians may be interested in a comparison of the two marginal distributions of row and column variables, say X and Y. One of such analyses is inferring whether X tends to be stochastically less than Y or vice versa. We are especially interested in applying the extension of the MH model, for example, the marginal cumulative logistic (ML) model (McCullagh, 1977; Agresti, 1984, p.205) based on the logit transform. The ML model states that one marginal distribution is a location shift of the other marginal distribution on a logistic scale. If the ML model fits the data poorly, we are then interested in other extension of the MH model based on the complementary log-log transform rather than logit transform.

Miyamoto, Niibe & Tomizawa (2005) gave the theorem that the MH model holds if and only if the ML model and the equality of row and column marginal means hold simultaneously. We refer to such relation as a decomposition of model (i.e., the MH model is decomposed into the ML model and the equality of row and column marginal means). Also, see Tahata & Tomizawa (2008) and Kurakami, Tahata & Tomizawa (2013) for the decompositions of the MH model. We are interested in whether the decomposition with the ML model replaced by the proposed model holds or not. When the MH model fits the data poorly, it may be useful for seeing the reason for the poor fit of it.

In this paper, Section 2 proposes a new model which is an extension of the MH model based on the complementary log-log transform. Section 3 gives the decomposition of the MH model using the proposed model. Section 4 refers to the goodness-of-fit test. Section 5 analyzes the father's and his son's occupational mobility data in Britain. We show that the new model and decomposition are useful for inferring relationships between marginal distributions with the example.

2. Models

For an $r \times r$ square contingency table with ordered categories, let p_{ij} denote the probability that an observation will fall in the ith row and jth column of the table for $i = 1, \ldots, r;\ j = 1, \ldots, r$. The MH model is defined by

$$p_{i\cdot} = p_{\cdot i} \quad (i = 1, \ldots, r),$$

where $p_{i\cdot} = \sum_{t=1}^{r} p_{it}$ and $p_{\cdot i} = \sum_{s=1}^{r} p_{si}$ (Stuart, 1955; Tahata & Tomizawa, 2014). This model indicates the structure that satisfies the identity of marginal distributions of row and column. Let F_i^X and F_i^Y denote the marginal cumulative probability of X and Y, respectively; namely $F_i^X = \sum_{s=1}^{i} p_{s\cdot}$ and $F_i^Y = \sum_{t=1}^{i} p_{\cdot t}$ for $i = 1, \ldots, r-1$. The MH model may also be expressed as

$$F_i^X = F_i^Y \quad (i = 1, \ldots, r-1).$$

Let L_i^X and L_i^Y denote the marginal cumulative logit transforms of X and Y, respectively; namely

$$L_i^X = \log\left(\frac{F_i^X}{1-F_i^X}\right), \quad L_i^Y = \log\left(\frac{F_i^Y}{1-F_i^Y}\right) \quad (i = 1, \ldots, r-1).$$

The MH model may further be expressed as

$$L_i^X = L_i^Y \quad (i = 1, \ldots, r-1).$$

The ML model (McCullagh, 1977) is defined by

$$L_i^X = L_i^Y + \delta \quad (i = 1, \ldots, r-1),$$

where the parameter δ is unspecified. The ML model is one of the extensions of the MH model. This model indicates that the odds that X is i or below instead of $i+1$ or above, is $\exp(\delta)$ times higher than the odds that Y is i or below instead of $i+1$ or above, for $i = 1, \ldots, r-1$. Therefore this model states one marginal distribution is a location shift of the other marginal distribution on a logistic scale.

Let C_i^X and C_i^Y denote the marginal cumulative complementary log-log transforms of X and Y, respectively; namely

$$C_i^X = \log\left(-\log\left(1-F_i^X\right)\right), \quad C_i^Y = \log\left(-\log\left(1-F_i^Y\right)\right) \quad (i = 1, \ldots, r-1).$$

The MH model may be expressed as

$$C_i^X = C_i^Y \quad (i = 1, \ldots, r-1).$$

We shall consider now the marginal cumulative complementary log-log (MCL) model which is defined by

$$C_i^X = C_i^Y + \log\Delta \quad (i = 1, \ldots, r-1),$$

where Δ is unspecified. This model indicates that the probability that X is $i+1$ or above, is equal to the probability that Y is $i+1$ or above to the power of Δ, for $i = 1, \ldots, r-1$. Thus this model states one marginal distribution is a location shift of the other marginal distribution on a complementary log-log scale. Note that if $\Delta = 1$, then we have the MH model. We see, under the MCL model, $\Delta > 1$ is equivalent to $F_i^X > F_i^Y$ and $\Delta < 1$ is equivalent to $F_i^X < F_i^Y$. Therefore the parameter Δ in the MCL model reflects the degree of inhomogeneity between $\{F_i^X\}$ and $\{F_i^Y\}$.

3. Decomposition

Consider the specified scores $\{u_k\}$ may be assigned to both rows and columns satisfying $u_1 \le u_2 \le \cdots \le u_r$ or $u_1 \ge u_2 \ge \cdots \ge u_r$, where at least one strict inequality holds. Using the function $g(k)$ which is $g(k) = u_k$ for $k = 1, \ldots, r$, consider the marginal mean equality (ME) model defined by

$$E(g(X)) = E(g(Y)),$$

where $E(g(X)) = \sum_{i=1}^{r} g(i)p_{i\cdot}$ and $E(g(Y)) = \sum_{i=1}^{r} g(i)p_{\cdot i}$.

We now obtain the following theorem.

Theorem 1. *The MH model holds if and only if both the MCL and ME models hold.*

proof. If the MH model holds, then the MCL and ME models hold. We assume that both the MCL and ME models hold, and then we show that the MH model holds. For $u_1 \le u_2 \le \cdots \le u_r$ (or $u_1 \ge u_2 \ge \cdots \ge u_r$), we have

$$E(g(X)) = \sum_{i=1}^{r} g(i)p_{i\cdot} = g(1) + \sum_{k=1}^{r-1} d_k\left(1-F_k^X\right),$$

where

$$d_k = g(k+1) - g(k).$$

Similarly, we have

$$E(g(Y)) = g(1) + \sum_{k=1}^{r-1} d_k \left(1 - F_k^Y\right).$$

Since the ME and MCL models hold, we have

$$\sum_{k=1}^{r-1} d_k \left(1 - F_k^X\right) = \sum_{k=1}^{r-1} d_k \left(1 - F_k^Y\right), \tag{1}$$

and

$$\sum_{k=1}^{r-1} d_k \left(1 - F_k^X\right) = \sum_{k=1}^{r-1} d_k \left(1 - F_k^Y\right)^{\Delta}. \tag{2}$$

Equations (1) and (2) lead to

$$\sum_{k=1}^{r-1} d_k \left(1 - F_k^Y\right) = \sum_{k=1}^{r-1} d_k \left(1 - F_k^Y\right)^{\Delta}.$$

Thus we obtain $\Delta = 1$, i.e., the MH model holds because $d_k \geq 0$ (or $d_k \leq 0$) for all $k = 1, \ldots, r-1$, with at least one of the $\{d_k\}$ being not equal to zero. The proof is completed.

4. Goodness-of-fit Test

Let n_{ij} denote the observed frequency in the ith row and jth column of the $r \times r$ table with $n = \sum \sum n_{ij}$, and let m_{ij} denote the corresponding expected frequency for $i = 1, \ldots, r;\ j = 1, \ldots, r$. We assume that a multinomial distribution applies to the table. The maximum likelihood estimates (MLEs) of expected frequencies under each model can be obtained using the Newton-Raphson method in the log-likelihood equation (see Appendix for the log-likelihood equation). The likelihood ratio chi-squared statistic for testing the goodness-of-fit of model M is given by

$$G^2(\mathrm{M}) = 2 \sum_{i=1}^{r} \sum_{j=1}^{r} n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right),$$

where $\hat{m}_{ij}$ is the MLE of m_{ij} under the model. The numbers of degrees of freedom (df) of statistics for testing the goodness-of-fit of the MH, ML, MCL, and ME models are $r-1$, $r-2$, $r-2$, and 1, respectively. Consider two nested models, say M_1 and M_2, such that if model M_1 holds, then model M_2 holds. For testing the goodness-of-fit of model M_1 assuming that model M_2 holds, the conditional likelihood ratio statistic is given by $G^2(\mathrm{M}_1|\mathrm{M}_2) = G^2(\mathrm{M}_1) - G^2(\mathrm{M}_2)$. The number of df for the conditional test is the difference between the numbers of df for the models M_1 and M_2.

5. Example

Consider the data in Table 1, relating the father's and his son's occupational status categories for a British sample again. The smaller category number means the higher status. We analyze the data using the new model and decomposition of the MH model.

Table 2 gives the values of likelihood ratio statistic G^2 for testing the goodness-of-fit of models. We set $u_k = k$ for $k = 1, \ldots, 5$. The MH, ML and ME models fit the data poorly (G^2(MH) = 32.80 with 4 df; G^2(ML) = 9.75 with 3 df; G^2(ME) = 20.28 with 1 df). The MCL model fits the data well (G^2(MCL) = 4.26 with 3 df). Using Theorem 1 which is the decomposition of the MH model into the MCL and ME models, we shall consider the reason why the MH model fits the data poorly. According to Theorem 1 and Table 2, the poor fit of the MH model is caused by the influence of the lack of structure of the ME model rather than the MCL model. Note that, using the decomposition of the MH model into the ML and ME models, it is difficult to consider the reason for the poor fit of the MH model because both the ML and ME models fit the data poorly.

Since the MCL model which is implied by the MH model fits well, we can test the goodness-of-fit of the MH model under the assumption that the MCL model holds, i.e., the hypothesis that $\Delta = 1$ under the assumption. The difference between the G^2 values for the MH and MCL models is G^2(MH|MCL) = G^2(MH) − G^2(MCL) = 28.54 with $4 - 3 = 1$ df, and thus the hypothesis that $\Delta = 1$ is rejected at the 0.05 significance level. It shows strong evidence of $\Delta \neq 1$ in the MCL model. Therefore the MCL model is preferable to the MH model for the data. Under the MCL model, the MLE of Δ is $\hat{\Delta} = 1.13$.

Namely, under the MCL model, the probability that the status category for father in a pair is $i+1$ or above, is estimated to be equal to the probability that the status category for son in the pair is $i+1$ or above to the power of 1.13, for $i = 1, \ldots, 4$. Since $\hat{\Delta} > 1$, under the MCL model, $\hat{F}_i^X > \hat{F}_i^Y$, where $\hat{F}_i^X$ and $\hat{F}_i^Y$ are MLEs of the marginal cumulative probabilities of X and Y for $i = 1, \ldots, 4$. Therefore the distribution of the status category for the son tends to be stochastically higher than that for his father.

Acknowledgements

The authors would like to thank two referees and the editor for their helpful comments.

References

Agresti, A. (1984). *Analysis of ordinal categorical data*. Wiley, New York. https://doi.org/10.1002/bimj.4710290113

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. MIT Press, Cambridge. https://doi.org/10.1007/978-0-387-72806-3

Kurakami, H., Tahata, K., & Tomizawa, S. (2013). Generalized marginal cumulative logistic model for multi-way contingency tables. *SUT Journal of Mathematics, 49*, 19–32. https://doi.org/10.20604/00000925

McCullagh, P. (1977). A logistic model for paired comparisons with ordered categorical data. *Biometrika, 64*, 449–453. https://doi.org/10.1093/biomet/64.3.449

Miyamoto, N., Niibe, K., & Tomizawa, S. (2005). Decompositions of marginal homogeneity model using cumulative logistic models for square contingency tables with ordered categories. *Austrian Journal of Statistics, 34*, 361–373. https://doi.org/10.17713/ajs.v34i4.424

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika, 42*, 412–416. https://doi.org/10.1093/biomet/42.3-4.412

Tahata, K., & Tomizawa, S. (2008). Generalized marginal homogeneity model and its relation to marginal equimoments for square contingency tables with ordered categories. *Advances in Data Analysis and Classification, 2*, 295–311. https://doi.org/10.1007/s11634-008-0028-1

Tahata, K., & Tomizawa, S. (2014). Symmetry and asymmetry models and decompositions of models for contingency tables. *SUT Journal of Mathematics, 50*, 131–165. https://doi.org/10.20604/00000822

Table 1. Occupational status for British father-son pairs (Bishop *et al.*, 1975, p.100)

	Son's status					
Father's status	1	2	3	4	5	Total
1	50	45	8	18	8	129
	(50.25)	(40.88)	(7.83)	(17.62)	(7.65)	(124.24)
2	28	174	84	154	55	495
	(31.07)	(172.82)	(90.59)	(166.05)	(57.80)	(518.33)
3	11	78	110	223	96	518
	(11.30)	(72.29)	(110.07)	(223.10)	(93.79)	(510.56)
4	14	150	185	714	447	1510
	(14.39)	(139.05)	(185.16)	(714.46)	(436.78)	(1489.84)
5	3	42	72	320	411	848
	(3.16)	(39.86)	(73.91)	(328.43)	(411.67)	(857.03)
Total	106	489	459	1429	1017	3500
	(110.17)	(464.90)	(467.57)	(1449.66)	(1007.70)	(3500.00)

Note: The parenthesized values are the MLEs of expected frequencies under the MCL model.

Table 2. Likelihood ratio chi-square values G^2 for models applied to the data in Table 1

Models	df	G^2
MH	4	32.80*
ML	3	9.75*
MCL	3	4.26
ME	1	20.28*

Note: u_k for the ME model is integer score. * means significant at the 0.05 level.

Appendix

We consider the MLEs of the expected frequencies $\{m_{ij}\}$ under the MCL model. Those under the MH, ML and ME models can be obtained in the similar manner, although those are omitted here.

To obtain MLEs under the MCL model, we must maximize the Lagrangian

$$L = \sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij}\log p_{ij} - \lambda\left(\sum_{i=1}^{r}\sum_{j=1}^{r} p_{ij} - 1\right) - \sum_{i=1}^{r-1}\mu_i\left(\log\left(1-F_i^X\right) - \Delta\log\left(1-F_i^Y\right)\right)$$

with respect to $\{p_{ij}\}$, λ, $\{\mu_i\}$, and Δ. Setting the partial derivatives of L equal to zero, we obtain the equations

$$p_{ij} = n_{ij}\left\{n + \sum_{k=1}^{r-1}\mu_k\left(\frac{F_k^X - I(i\le k)}{1-F_k^X} - \frac{\Delta\left(F_k^Y - I(j\le k)\right)}{1-F_k^Y}\right)\right\}^{-1} \quad (i=1,\ldots,r;\ j=1,\ldots,r),$$

as well as

$$1 - F_i^X = \left(1-F_i^Y\right)^{\Delta} \quad (i=1,\ldots,r-1),$$

and

$$\sum_{i=1}^{r-1}\mu_i\log\left(1-F_i^Y\right) = 0,$$

where $I(\cdot)$ is the indicator function. Using the Newton-Raphson method, we can solve the equations with respect to $\{p_{ij}\}$, $\{\mu_i\}$ and Δ. Then we can obtain the MLEs of $\{m_{ij}\}$ and Δ under the MCL model.

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Modeling Trend in Telecommunication in Sri Lanka:

A Case study on Internet and Cellular Connections

K. A. N. K. Karunarathna[1], S. Brindha[1] & P. Paramadevan[1]

[1] Department of Mathematics, Faculty of Science, Eastern University, Sri Lanka

Correspondence: K. A. N. K. Karunarathna, Department of Mathematics, Faculty of Science, Eastern University, Sri Lanka.

Received: April 4, 2018 Accepted: April 18, 2018 Online Published: May 28, 2018

doi:10.5539/ijsp.v7n4p32 URL: https://doi.org/10.5539/ijsp.v7n4p32

Abstract

The telecommunication is one of the modes of communication, in which most investments are made. It consists of internet, mobile phones, wired and wireless fixed phones, fax, televisions, radio and some other. Among them, demand for internet and cellular phones rapidly increases. For a smooth function of this business, knowledge on demand is much important. Effective forecasts help a business to manage its supply efficiently. This study aimed to find out an accurate mechanism for prediction of demand for internet conections and cellualr phone collections.

Based on the secondary data available in central bank reports from 1996 to 2016, several statistical forecasting models were evaluated for an accurate prediction. There can be seen an increasing demand for both internet and cellular phone connections. Number of internet connections has gone up from 4 110 to 4 921 000, while the usage of cellular phones has developed from 71 228 to 26 228 000 during this period. Rapid growth in internet usage has happened after 2009, while after year 2003, usage of cellular phone has increased rapidly. With compared to models fitted for original form of data, models for log transformed data show better performances. The best performance in prediction of internet connection was given by ARIMA (1,1,1) model fitted for log transformed data, meanwhile ARIMA (0,1,2) model fitted for log transformed data showed the best fit for series of cellular connections. Double exponential smoothing models also show better fit for both series.

Keywords: ARIMA, Cellular phones, Demand, Forecasting, Internet, Telecommunication

1. Introduction

Communication is the process of transferring data and information from a source to a destination. Voice, body language and signs are the simplest modes of the communication. This information may be in the forms of audio, video, graphics, writings, images, gestures, signs and many more.

Advancement in technologies has changed modes of communication over last fifty years, which gave rise to the telecommunication. At present, people communicate through emails, faxes, mobile phones, texting services, video conferences, video chat rooms and social media and many more to evolve in upcoming years. This is known as telecommunication and it is one of the most important and rapid growing industries at the present era. The most significant telecommunication aspects are the internet, satellites and the cellular phones. These modes of communication have increased speed of transferring and exchanging data to a greater distant with a low cost effectively.

The very common form of telecommunication service is the phone service, which is done on either a wired or wireless form. The internet, television, and networking for businesses and domestic purposes are among the other services. These services may not be available in all areas or from all companies. The pricing points for different services vary widely and may be different for residences and businesses. These options are now expanded to wireless connections, while some companies offer both wireless and landline services together. Some service providers are offering television now, with a higher bandwidth speeds available through an improved infrastructure such as fiber optics. Optical fiber has revolutionized the modern telecommunication industry. It helps in transferring information to much greater distance as it provides higher bandwidth with little or no loss in the transmission medium.

1.1 Cellular Usage

Cellular phone was invented in the early 1970s, which was not much noticed. At that time, the use of cell phone was limited to certain areas and the cost was not affordable by everyone. As the technology rises, the phone came down in size,

price and weight, which took the attention of the entire nation. The cellphone changed our lifestyles and took place next to our wallets in the pocket. It gives us the instant and constant communication with the mobility we desperately needed. Modern cellphones are designed in such a way it contains all the personalized device which is owned by an individual which includes camera, mp3/mp4 player, games, document folders, etc. Mobile phone's size is getting bigger and bigger day by day with the high updated technologies in it with inbuilt batteries. Even though the price is higher, adults to teens are buying them to get the full benefit from it. As a result, they do not depend on a landline to communicate. Hence, the usage of landlines and public phone booth is declining.

1.2 Internet Usage

Internet has bought a huge impact in our lives. Since it was found that it has brought information and knowledge on our fingertips. Internet has brought positive impact in our lives and has made it simple and easier than ever. Earlier in search of information, we have to travel all the way to library or get suggestions from the elders, now the use of libraries have reduced to a greater extent due to introduction of the internet services. We are able to access large and excess data in just one click. It helps in utilizing our time in a productive manner. The most important use of internet is that it gives information and education. It provides with various websites and various blogs that give informative and useful content which helps the students in studies. It helps the people to learn various things and people get knowledge which they implement in their daily life. It helps in communication with the people easily and faster than before. We are able to send e-mails, video chatting, texting, watch movies and dramas, shop on e-shopping websites so on. ICT-Information and communication Technology has given wings to empower the use of technology related activities in the educational world. It is growing in a skyrocketing speed. ICT is used in daily life such as in education, banking, business and all the industrial uses. It helps us in e-learning, online banking, access books online, helps in presentation and researching and many more.

1.3 Objectives

Since the technology has improved and changed the telecommunication sector to a greater extend, it is important to identify the trend in the usage of these services. Therefore, this study takes the facts and figures of cellular phones and internet as the sub-indices of tele-communication sector into account for the time analysis of the data. Effective forecasts help a business manage its supply chain more economically and efficiently. Accurate predictions allow a business to manufacture and services more favorably because it has sufficient time to evaluate and plan. An accurate forecast enables a business owner to keep a lower inventory and thus reducing costs and wastages.

1.4 Previous Works

According to David (2011), "Forecasts are educated assumptions about future trends and events". Demir and Ozsoy (2014), have stated that forecasting is a complicated process as the factors such as innovation in technology, changes in culture and social values, unstable economic conditions, new product, stronger competitors, improved services, etc. There are different models for forecasting and their accuracies are depending on the situations and data considered.

Dhanushka (2013) has examines the growth of the telecommunication sector in Sri Lanka's by using annual time series. This study consisted of bivariate and multivariate co-integration approach to establish the long run equilibrium relationship and causality testing to detect the direction of this relationship. According to author, this study is the first of its kind to use annual secondary data to examine the long run relationship between telecommunications sector and service sector in Sri Lanka. Also one-way link between telecommunications sector growth and service sector growth was established through causality test. The sample has confirmed that research hypothesis is positive for the collected data. Thus, it has been concluded that increase in telecommunications sector growth increases the long run service sector growth.

Chakaraborty, and Nandi (2003) have examines the relationship between the level of telecommunications infrastructure (measured by telephone mainlines per capita or tele density rate) and economic growth by exploiting a panel co-integration framework. Almost all of these studies have documented a positive correlation between tele density rates and a variety of indices of economic growth. The conclusions are based on simple correlation coefficient and regression analysis. Given the unit root characteristics of time series variables in general, results based on regression analysis, as pointed out by many, are subject to spurious correlation. In addition, the simple regression coefficients fail to establish the causal relationship and its direction between the variables of interest. The study also examines the relationship between the two referred variables for a panel of 12 Asian developing countries that vary in terms of stages of development. Canning, and Pedroni (2004), TatyanaPalei (2014), Farhadi, Ismail, Fooladi (2012), Lee, and Alford (2017) also have shown the impact of telecommunication on economic growth in different regions.

It seems that studies on modeling usage of cellular and internet connections in literature are lacking. But, studies on modeling some other responses are available. Fatai *et. al.* (2003), has used Engle-Granger's error correction model(ECM), and the autoregressive distributed lag regression approaches(ARDL) to model the demand for electricity in New Zealand.

Abraham and Nath (2001) have used Box-Jenkins autoregressive integrated moving average (ARIMA) approach in modeling electricity in the state of Victoria, Australia. Monthly electricity consumption in Pakistan has been analyzed by Yasmeen and Sharif (2015), by using both linear and nonlinear modeling techniques including ARIMA, Seasonal ARIMA (SARIMA) and ARCH/GARCH models. This study evaluated some of these models and some other for modeling usage of internet and cellular phones.

2. Method

2.1 Data Collection

Secondary data, available in Sri Lankan central bank's annual reports, were used for this study. Number of internet connections and cellular connections are in use were recorded with year for the period from year 1996 to year 2016. Data in required form were not available for early period.

2.2 Statistical Models

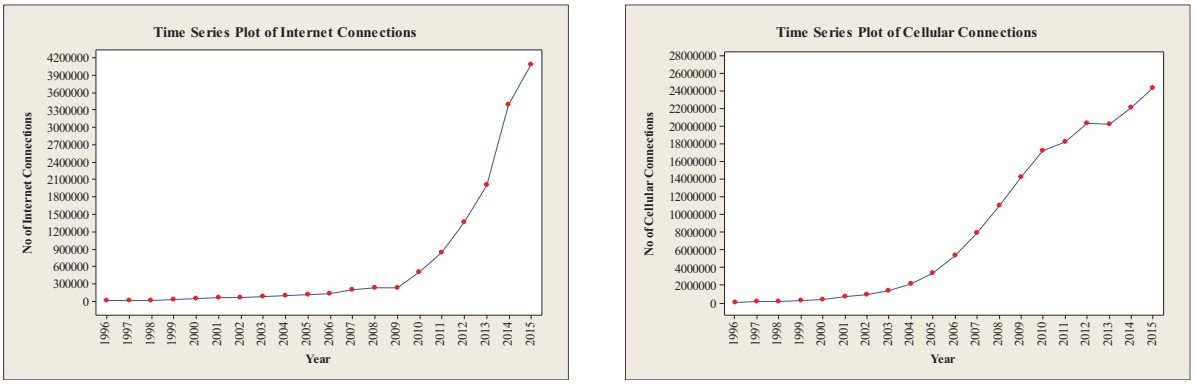
There are several statistical models that can be used to explain trends in a series. Among them, Single exponential smoothing model (SESM), double exponential smoothing model (DESM), growth curve model(GCM), quadratic trend model(QTM), auto regressive model (ARM), moving average model(MAM), auto regressive moving average model(ARMAM), auto regressive integrated moving average model(ARIMAM) were evaluated to model the number of internet connections, and cellular phone connections. These models were tested for both original form and transformed form of data. As the transformation, natural logarithm was used. In addition to the above models, linear trend models (LTM) were also evaluated only for log transformed data.

In fitting Box Jenkins AR, MA, ARMA, and ARIMA models, stationarity of series was tested with the help of autocorrelation and partial autocorrelation functions. When the series was not stationary, by taking the first differences, series was made stationary.

As the accuracy measures of forecast of each model, mean absolute percentage error(MAPE), mean absolute deviation(MAD), mean squared deviation(MSD) were used in case of SESM, DESM, GCM, QTM, LTM while the sum of square errors(SS), mean square error(MSE) were used additionally in case of MA, AR, ARMA, and ARIMA models. For the analysis, 14th version of the statistical package, Minitab, was used.

3. Results

Trend in usage of internet and cellular phones is exhibited in Figure 1(a) and Figure 1(b) respectively during the period from1996 to 2015. Both series of internet and cellular connections show increasing patterns during the period considered. Number of internet connections has gradually increased during this period. However, two different phases can be observed in this pattern. During the period from 1996 to 2009, number of internet connections has increased almost linearly from 4 110 to 240 000 with a rate of 18 145 per year. Number of internet connection has developed from 240 000 to 4 091 000 in the period from 2009 to 2015. In this period, average increment in number of internet connections per year is about 641 833.



(a) Internet (b) Cellular

Figure 1. Trend in series of Internet and Cellular connections

Numbers of cellular connections also have increased year by year during this period except in year 2013. Number of cellular connections in 2013 has decreased than the preceding year. In the series of number of cellular connection also,

two phases can be observed depending on the trend. From 1996 to 2002, number of cellular connections has developed from 71 228 to 931 580, on average, at a rate of 102 815 per year, while number of cellular connections has developed from 931 580 to 24 385 000 in the period from 2002 to 2015 with an average rate of 1 804 109 per year. However, it is not a linear increment.

Details, related to the best model selected from SESM, DESM, GCM, QTM for internet connections, are given in Table 1 below, while details of fitted ARIMA models given in Table 2. Even though, a model from each SESM, DESM, GCM, QTM, MAM, ARM could be fitted, an ARIMA model could not be found for the original form of data.

Table 1. Models for series of Internet connections

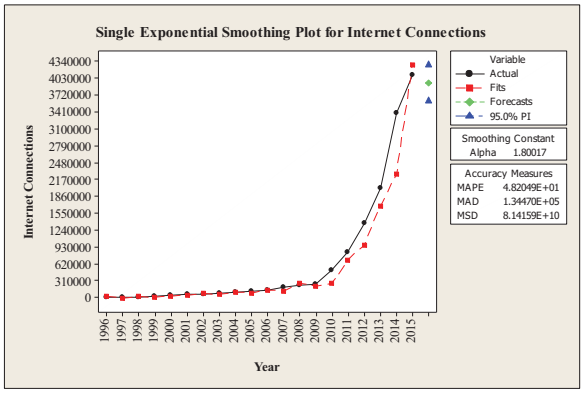
Series	Model	Alpha	Gamma	MAPE	MAD	MSD
Original Series	SESM	1.8001	-	4.82E+01	1.34E+05	8.14E+10
	DESM	0.1333	8.9569	1.38E+02	1.03E+05	2.97E+10
	GCM	-	-	3.00E+01	1.63E+05	1.16E+11
	QTM	-	-	9.39E+02	3.44E+05	1.57E+11
		SS	**MS**	**MAPE**	**MAD**	**MSD**
	ARIMA (0,1,1)	1.63E+12	9.04E+10	3.05E+01	1.14E+06	8.56E+10
	ARIMA (1,1,0)	1.07E+12	5.93E+10	1.54E+01	1.18E+05	5.62E+10

Log series	Model	Alpha	Gamma	MAPE	MAD	MSD
	SESM	1.8995	-	2.3033	0.2664	0.1000
	DESM	0.8725	0.6144	1.8238	0.2035	0.0672
	LTM	-	-	2.4616	0.2809	0.1091
	GCM	-	-	2.2815	0.2463	0.1057
	QTM	-	-	2.3635	0.2623	0.1047
		SS	**MS**	**MAPE**	**MAD**	**MSD**
	ARIMA(0,1,1)	1.7199	0.0956	2.1462	0.2573	0.0905
	ARIMA(1,1,0)	1.1270	0.0626	1.5564	0.1872	0.0593
	ARIMA(1,1,1)	1.0211	0.0601	1.5231	0.1797	0.0537

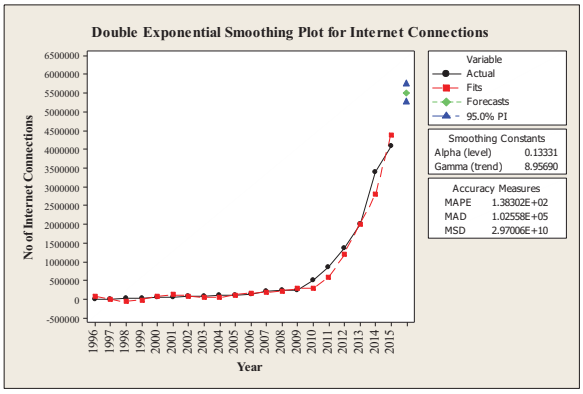
Table 2. ARIMA Models for series of Internet connections

Series	Model	Coef	SE	P-value
Original Series	ARIMA (0,1,1)	-0.8002	0.255	0.0060
	ARIMA (1,1,0)	0.8917	0.1465	0.0000
Log Series	ARIMA(0,1,1)	-0.8995	0.1096	0.0000
	ARIMA(1,1,0)	0.9301	0.1220	0.0000
	ARIMA(1,1,1)	0.9981	0.0613	0.0000
		0.5181	0.2141	0.0270

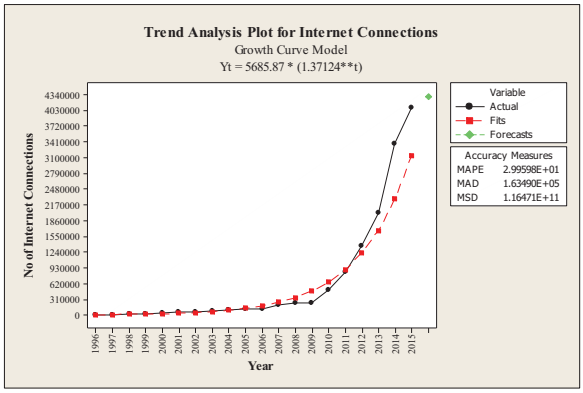
SESM and DESM are depending on some parameters called “alpha” and “gamma” and optimal values of them are given in Table 1. Among the models SESM, DESM, GCM, and QTM, GCM shows the minimum MAPE while QTM shows the highest. However, DESM and QTM show the minimum and the maximum of MAD respectively. DESM shows the least MSD. Models ARIMA (0,1,1) and ARIMA (1,1,0) only could be fitted for the series of the internet connections. Out of these two models, ARIMA (1,1,0) model shows the minimum for SS, MS, MAPE, MAD, and MSD.



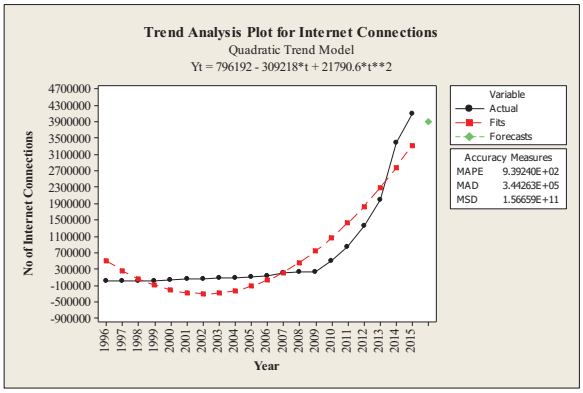
(a) SESM



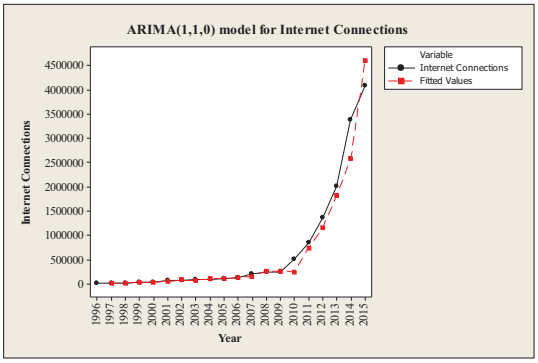
(b) DESM



(c) GCM



(d) QTM

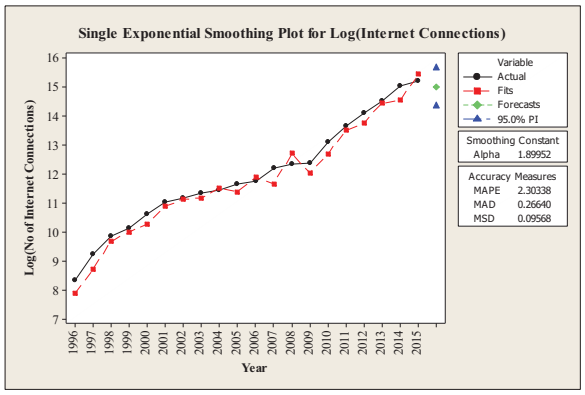


(e) ARIMA (1,1,0)

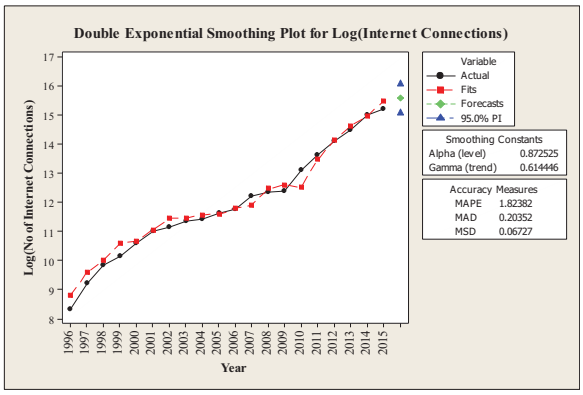
Figure 2. Plot of observed and fitted values of each model for internet connections

Fitted values with each model alone with the observed values are plotted in Figure 2. Those plots show how far fitted values are closer to the observed values under each model. According to them, it can be seen that DESM and ARIMA (1,1,0) give better estimates for observed values.

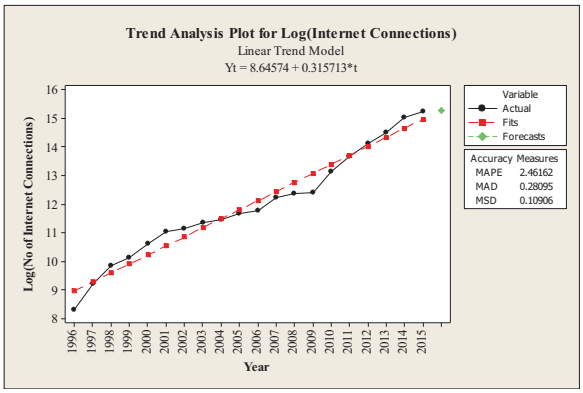
Details of models that fitted for log transformed number of internet connection are also given in the Table 1. DESM shows the minimum MAPE (1.82382) while the other models are having a higher almost the same MAPE. In case of MAD, DESM shows the minimum while LTM shows the highest. MSD of DESM shows the least while other models show little higher similar values.



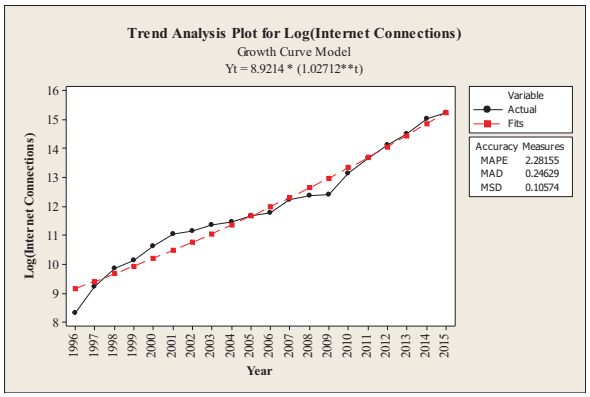
(a) SESM



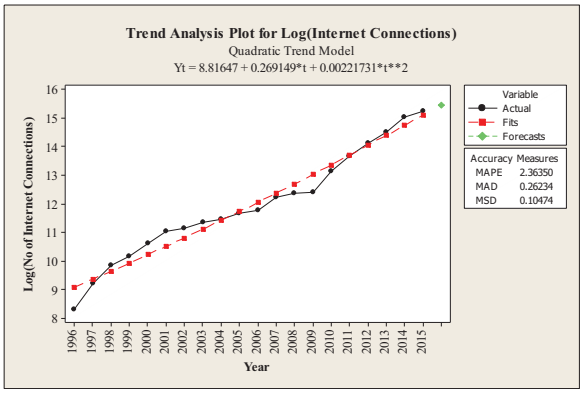
(b) DESM



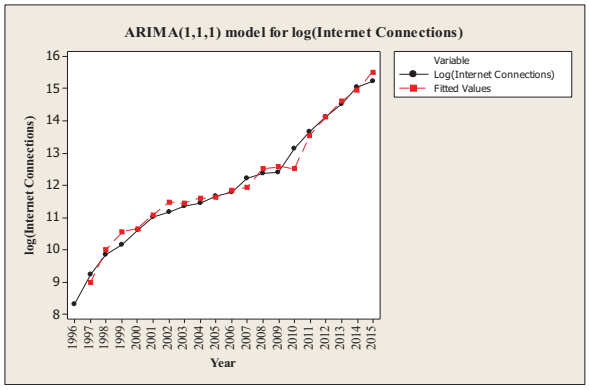
(c) LTM



(d) GCM



(e) QTM



(f) ARIMA(1,1,1)

Figure 3. Plot of observed and fitted values of each model for log(internet connections)

For series of the first differences of log transformed number of internet connections, MA (1) [ARIMA (0,1,1)], AR (1) [ARIMA (1,1,0)] and ARIMA (1,1,1) models could be fitted without a constant. Details are in the Table 1 above. With compared to ARIMA (0,1,1) and ARIMA (1,1,0) models, ARIMA (1,1,1) has given the lowest SS, MS, MAPE, MAD, and MSD. Plots of observed and fitted values obtained from each model are exhibited in Figure 3. It is clear that DESM and ARIMA (1,1,1) gives better prediction with compared to the other models.

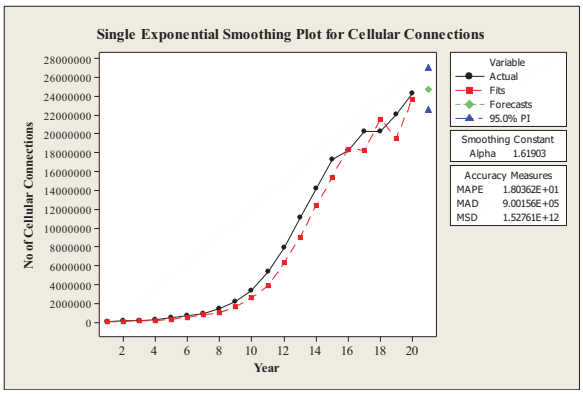
Table 3. Models for series of Cellular connections

Series	Model	Alpha	Gamma	MAPE	MAD	MSD
Original Series	SESM	1.6190	-	1.80E+01	9.00E+05	1.53E+12
	DESM	0.1519	11.5293	3.63E+01	4.11E+05	3.53E+11
	GCM	-	-	3.97E+01	4.80E+06	9.71E+13
	QTM	-	-	1.29E+02	1.29E+06	2.31E+12
		SS	**MS**	**MAPE**	**MAD**	**MSD**
	ARIMA(0,1,2)	1.58E+13	9.35E+11	4.12E+01	7.43E+05	8.36196E+11
	ARIMA(1,1,0)	1.35E+13	7.52E+11	9.76E+00	5.89E+05	7.12044E+11
Log series	**Model**	**Alpha**	**Gamma**	**MAPE**	**MAD**	**MSD**
	SESM	1.9126	-	1.2686	0.1796	0.0400
	DESM	0.9318	1.0171	0.4584	0.0671	0.0055
	LTM	-	-	2.4510	0.3660	0.1864
	GCM	-	-	3.0596	0.4589	0.2943
	QTM	-	-	-	-	-
		SS	**MS**	**MAPE**	**MAD**	**MSD**
	ARIMA(0,1,2)	0.2621	0.0154	0.6843	0.0991	0.0138

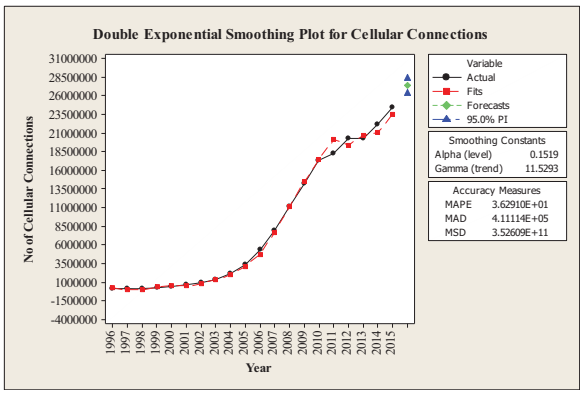
Same models were tested for the series of number of cellular connections. Both series of original form of data and transformed data were modeled. Details of the selected models are given in the Table 3 and Table 4. A SESM with an alpha value 1.619 could explain the series better than SESM with other alpha. Among DESM with different alpha and gamma, a model with alpha of 0.1519 and gamma of 11.5293 could be selected as the best DESM for series of original data of cellular connections. Among the models SESM, DESM, GCM, QTM fitted for the original form of cellular connections, the minimum MAPE is given by SESM while DESM gives the minimum of MAD and MSD. Further, ARIMA (0,1,2) and ARIMA (1,1,0) models could be fitted as the best models from each type for this series. According to SS, MS, MAPE, MAD, and MSD, ARIMA (1,1,0) model is better than ARIMA (0,1,2) for cellular connections.

Among the models fitted for log transformed cellular connections, DESM gives the minimum of MAPE (0.4584) while it gives the minimum for MAD also. In case of MSD also, DESM shows the least. No any AR, MA or ARMA models could be fitted for the log transformed data of cellular connections.

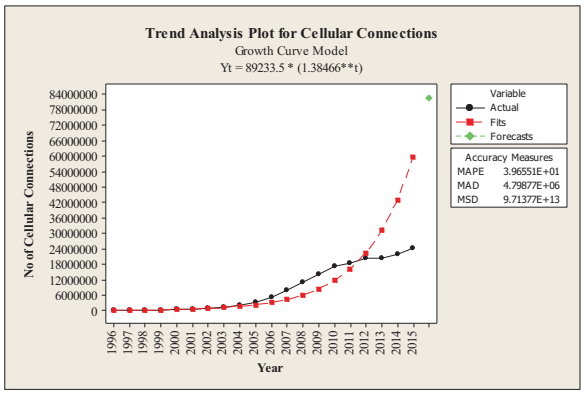
However, only ARIMA (0,1,2) model could be fitted for the series of log transformed data. It shows a SS of 0.2621 and MS of 0.0154, while it gives low values for MAPE, MAD, and MSD.



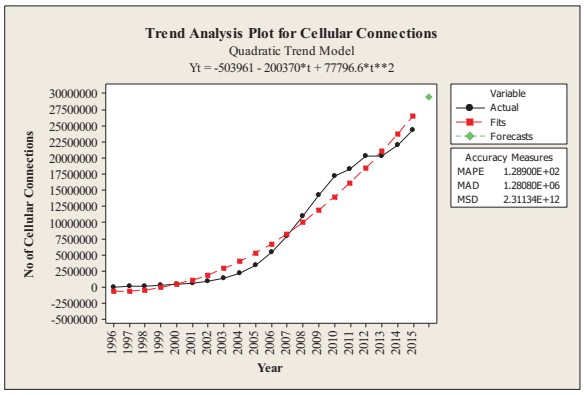
(a) SESM



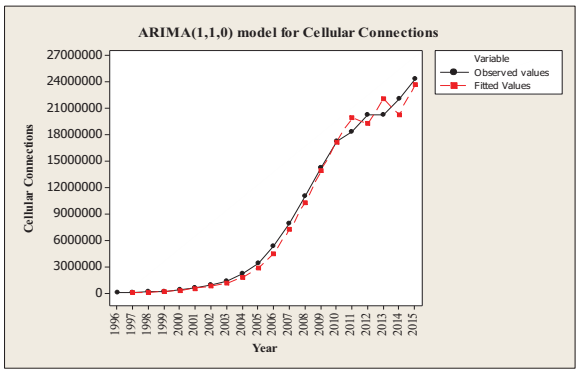
(b) DESM



(c) GCM



(d) QTM



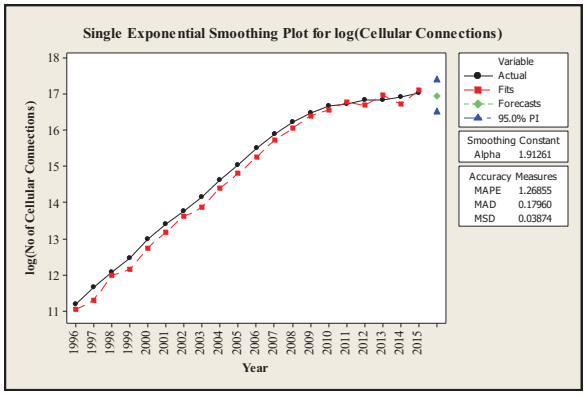
(e) ARIMA (1,1,0)

Figure 4. Plot of observed and fitted values of each models for cellular connections

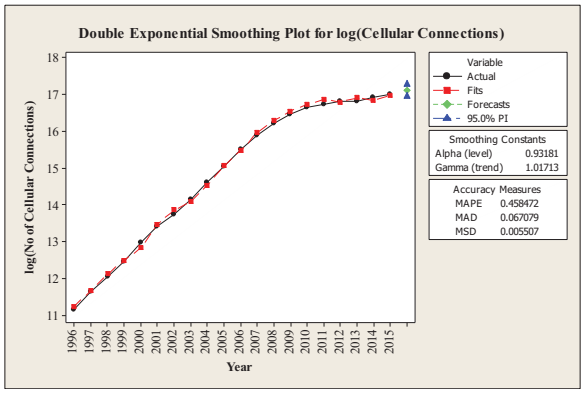
Table 4. ARIMA models for series of Cellular connections

Series	Model	Coef.	SE	P-value
Original series	ARIMA(0,1,2)	-0.5228	0.2332	0.0390
		-0.8786	0.2332	0.0020
	ARIMA(1,1,0)	0.9125	0.1223	0.0000
Log series	ARIMA(0,1,2)	-1.5229	0.1762	0.0000
		-0.9361	0.1710	0.0000

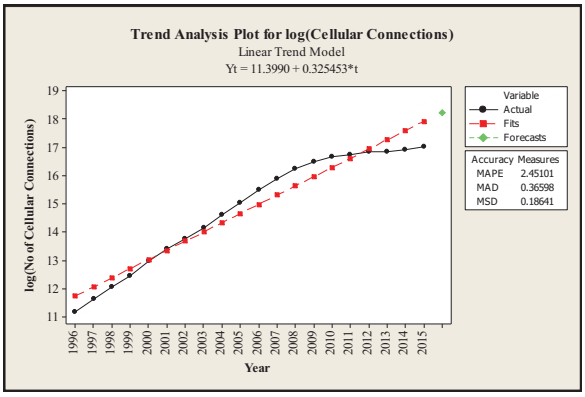
Plot of fitted values from selected models and observed values of log transformed cellular connections are given in Figure 5. It is clear that DESM and ARIMA (0,1,2) fit data well with compared to other models.



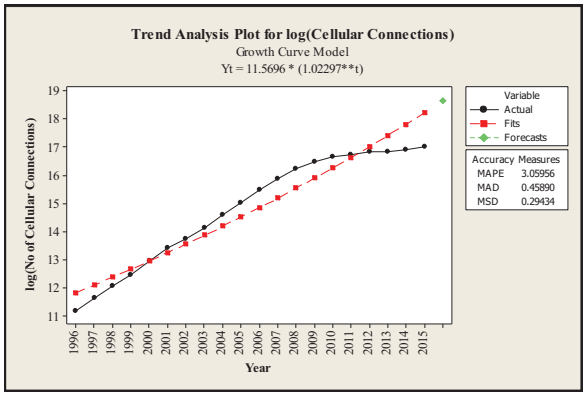
(a) SESM



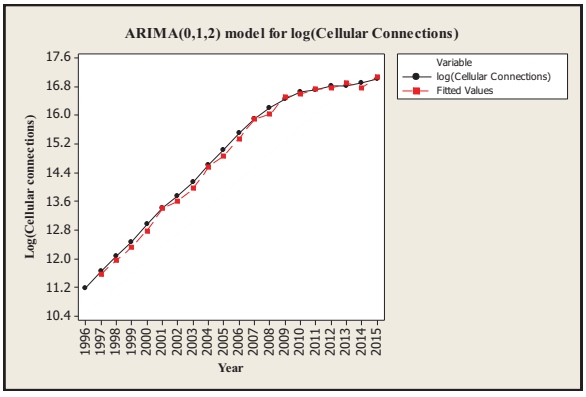
(b) DESM



(c) LTM



(d) GCM



(e) ARIMA (0,1,2)

Figure 5. Plot of observed and fitted values of each model for log (cellular connections)

Based on the most suitable models, predictions were made for year 2016. Forecasted values from each selected model, forecasting error as a percentage and 95% confidence intervals are given in the Table 5.

Table 5. Forecasted values of selected models

Series	Selected Model	Predicted value	Prediction Error	(%)	95% Confidence Interval of Predicted Value
Internet	DESM	5 495 397	-11.6723		(5 244 135, 5 746 659)
	ARIMA(0,1,2)	4 221 454	14.2155		(3 718 076, 4 724 833)
log(Internet)	DESM	5 877 252	-19.4321		(3 569 728, 9 676 395)
	ARIMA(1,1,1)	5 726 413	-16.3669		(3 541 638, 9 258 012)
Cellular	DESM	27 394 084	-4.4459		(26 386 874, 2 840 129)
	ARIMA(0,1,2)	26 141 188	0.3309		(24 246 010, 2 836 366)
log(Cellular)	DESM	26 923 228	-2.6507		(22 844 021, 31 734 024)
	ARIMA(0,1,2)	25 610 167	2.3556		(20 075 284, 32 667 785)

According to the prediction error corresponding to year 2016, DESM gives the minimum error of prediction for series of internet, while ARIMA (1,1,1) model gives a higher error. In case of cellular connections, ARIMA (0,1,2) model fitted for original form of data gives the minimum while ARIMA (0,1,2) model fitted for log (cellular connections) gives relatively a lager error.

4. Discussion

This study aimed to find a suitable statistical model that fit well with number of internet connections and number of cellular connections. Several models were evaluated for original and log transformed data of these series. With compared to models fitted for original form of data of both internet and cellular connections, models fitted for log form of data performed well.

There is an increasing trend in usage of both internet and cellular phones during the period from 1996 to 2015. However, there is a rapid growth in usage of internet after year 2009 while a rapid development in cellular phone can be seen after years about 2003, 2004. Model ARIMA (1,1,1) fitted for log transformed data can fit the behavior of the series of internet connections and this model can be used for the predictions. In addition to this model, double exponential smoothing model, fitted for log-transformed data with alpha of 0.8725 and gamma of 0.6144, also can explain data well.

Fluctuations of series of cellular connections could be explained by using ARIMA (0,1,2) model with log transformation. As an alternative model, double exponential smoothing model with alpha of 0.9318 and gamma of 1.0171 also could be used for prediction with log-transformed data.

In selecting the best model, it is necessary to compare accuracy measures of each model fitted for data in different forms (original form and log form). Model that gives the minimum for those measures, is supposed to be the best. Models fitted for log-transformed data, produced small values for summary measures. Then, it was difficult to make comparisons with the summary measures of models fitted for original data, which are large. Parallel to the accuracy measures, prediction errors also should be taken into account in selecting a model. However, in this study, priority was given to the accuracy measures mentioned above because they are on averages and forecasting error in Table 5 is a just single value.

Since there are some similarities in trends of these two series, there seems to be a possibility for multivariate approaches such as multivariate regression, and vector autoregressive models. They are to be evaluated at the next step.

5. Conclusions

During this period concerned, usage of internet connections has increased from 4 110 to 4 921 000, while the usage of cellular phones has gone up from 71 228 to 26 228 000. After 2009, a significant growth in internet usage could be observed, while usage of cellular phone has increased rapidly after year 2003. Among all models considered, models fitted for log transformed data show better performances. ARIMA (1,1,1) model fitted for log transformed data showed the best performance in prediction of internet connection, while ARIMA (0,1,2) model fitted for log transformed data showed the best fit for series of cellular connections. Double exponential smoothing models also show better fit for both series.

References

Abraham, A., & Nath, B. A. (2001). Neuro-fuzzy approach for modeling electricity demand in Vitoria. *Applied Soft Computing, 1*, 127-138.

Canning, D., & Pedroni, P. (2004). The Effect of Infrastructure on Long Run Economic Growth. Harvard University.

Chakaraborty, C., & Nandi, B. (2003). Privatization, Telecommunications and Growth in Selected Asian Countries: An Econometric Analysis. *Communications and Strategies, 52*(4), 31–47.

David, F. R. (2011). Strategic management: Concepts and cases. New Jersey. Pearson Education.

Demir, A., & Ozsoy, S. (2014). Forecasting the monthly electricity demand of Georgia using competitive models and advises for the strategic planning. *International Journal of Academic Research in Economics and Management Sciences. 3*, 90-103.

Dhanushka, T. (2013). The impact of telecommunication growth on the service sector: a cointegration analysis. Journal of Management, 09(01),

Farhadi, M., Ismail, R., & Fooladi, M. (2012) Information and Communication Technology Use and Economic Growth. PLoS ONE 7(11): e48903. https://doi.org/10.1371/journal.pone.0048903

Fatai, K., Oxley, L., & Scrimgeour, F. G. (2003). Modelling and forecasting the Demand for electricity in New Zealand: A comparison of alternative approaches. *Energy Journal, 24*(1), 75-102.

Maryam, F., Rahmah, I., & Masood, F. (2012), Information and Communication Technology use and economic growth. PLoS ONE7(11), e48903. https://doi.org/10.1371/journal.pone.0048903

Paley, T. (2014, October). Assessing the Impact of Infrastructure on Economic Growth and Global Competitiveness. Paper presented at the second Global Conference on Business, Economics, Management and Tourism, Prague, Czech Republic. Retrieved from https://ac.els-cdn.com/S2212567115003226/1-s2.0-S2212567115003226-main.pdf?_tid=4fdc1c5a-584c-4f91-bff2-45ad3630e8e3&acdnat=1522138297_ 82b83d 3bce9a961ed2eff77b3090ddc8

Sang, L., & Mathew, A. (2017). The effect of information communication technology on stock market capitalization: A panel data analysis Telecommunications Infrastructure and Economic Growth: Evidence from Developing Countries, *Business and Economic Research, 7*(1).

Yasmeen, F., & Sharif, M. (2015). Functional time series forecasting of electricity consumption in Pakistan. *International Journal of Computer Application, 124*, 15-19.

Copyrights

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

New Bounds on Poisson Approximation for Random Sums of Independent Binomial Random Variables

Giang Truong Le[1]

[1] University of Finance - Marketing, Vietnam

Correspondence: Giang Truong Le , University of Finance - Marketing, 2/4 Tran Xuan Soan street, District 7, Ho Chi Minh city, Vietnam.

Received: April 10, 2018 Accepted: April 25, 2018 Online Published: June 19, 2018

doi:10.5539/ijsp.v7n4p43 URL: https://doi.org/10.5539/ijsp.v7n4p43

Abstract

In this paper, we use the Stein-Chen method to obtain new bounds on Poisson approximation for random sums of independent binomial random variables. Some results related to sums of independent binomial distributed random variables are also investigated. The received results in the present study are more general and sharper than some known results.

Keywords: Binomial random variable, Poisson approximation, Random sums, Stein-Chen method

1. Introduction

In recent times, Poisson approximation problem for random sums of discrete random variables has attracted the attention of mathematicians. Readers who are interested in this problem can refer to (Hung & Giang, 2016b), (Kongudomthrap & Chaidee, 2012), (Teerapabolarn, 2013), (Teerapabolarn, 2014b), (Vellaisamy & Upadhye, 2009) and (Yannaros, 1991) for more details. We need to recall some results concerning the bounds in Poisson approximation for random sums of discrete random variables.

Let $Z_1, Z_2, \ldots$ be a sequence of independent Bernoulli random variables, each with probability of success $P(Z_i = 1) = p_i = 1 - P(Z_i = 0), i = 1, 2, \ldots$, and let N be a positive integer-valued random variable and independent of Z_i's. Let U_{λ^*} be a Poisson random variable with mean λ^*, $V_N = \sum_{i=1}^{N} Z_i$, $\lambda^* = E\left(\lambda_N^*\right)$ and $\lambda_N^* = \sum_{i=1}^{N} p_i$. In 1991, Yannaros gave a uniform bound for the total variation distance between the distributions of V_N and U_{λ^*} as follows, see (Yannaros, 1991):

$$d_{TV}\left(V_N, U_{\lambda^*}\right) \leq E\left|\lambda_N^* - \lambda^*\right| + E\left(\frac{1 - e^{-\lambda_N^*}}{\lambda_N^*} \sum_{i=1}^{N} p_i^2\right). \tag{1}$$

Let $X_1, X_2, \ldots, X_n$ be n independently distributed binomial random variables, each with probabilities

$$P\left(X_i = k\right) = C_{r_i}^k p_i^k (1 - p_i)^{r_i - k},$$

where $p_i \in (0, 1); r_i = 1, 2, \ldots; i = 1, 2, \ldots n; k = 0, 1, \ldots r_i; C_{r_i}^k = \dfrac{r_i!}{k!\,(r_i - k)!}$.

Suppose that N is a positive integer-valued random variable and independent of X_i's. Let U_λ be a Poisson random variable with mean λ, $W_N = \sum_{i=1}^{N} X_i$, $\lambda_N = \sum_{i=1}^{N} r_i p_i$ and $\lambda = E\left(\lambda_N\right)$. In 2014, Teerapabolarn used the Stein-Chen method to obtain a uniform bound for the total variation distance between the distribution functions of W_N and U_λ as follows, see (Teerapabolarn, 2014a):

$$d_{TV}\left(W_N, U_\lambda\right) \leq E\left(\frac{1 - e^{-\lambda_N}}{\lambda_N} \sum_{i=1}^{N} r_i p_i^2\right) + \min\left\{1, \sqrt{\frac{2}{\lambda e}}\right\} E\left|\lambda_N - \lambda\right|. \tag{2}$$

This paper is organized as follows. The second section is a brief introduction to Stein-Chen method. In section 3, we give main results of this paper, and conclusions of this study are presented in the last section.

In addition, throughout this paper, d_{TV} is denoted the total variation distance, defined by

$$d_{TV}\left(X, Y\right) = \sup_A \left|P\left(X \in A\right) - P\left(Y \in A\right)\right|,$$

where $A \subseteq \mathbb{Z}_+ := \{0, 1, 2, \ldots\}$.

2. Preliminaries

The Stein-Chen method has been dealt with in detail in many articles (the reader is referred to (Chen, 1975) and (Barbour, Holst & Janson, 1992) for fuller development). The Stein-Chen method can be summarized as follows.

Let us denote by $F_{W_n}(A)$ the probability distribution function of a discrete random variable $W_n \in A$ and we will be denoted by $P_{\lambda_n}(A) = \sum_{k\in A} e^{-\lambda_n} \frac{{\lambda_n}^k}{k!}$ the Poisson distribution function ($\lambda_n > 0$), defined on the set $A \subseteq \mathbb{Z}_+$. The best known method for estimating

$$\Delta = \sup_x \left| F_{W_n}(A) - P_{\lambda_n}(A) \right|$$

is basing on the following arguments (see (Chen, 1975) for more details).

Assume that h is a bounded real-valued function defined on $\mathbb{Z}_+$ and

$$P_{\lambda_n}(h) = e^{-\lambda_n} \sum_{k=0}^{\infty} h(k) \frac{{\lambda_n}^k}{k!}.$$

Consider the function $f(.)$ which is a solution of the Stein's equation

$$\lambda_n f(w+1) - w f(w) = h(w) - P_{\lambda_n}(h). \tag{3}$$

Setting

$$h(w) = h_A(w) = \begin{cases} 1, & \text{if} \quad w \in A, \\ 0, & \text{if} \quad w \notin A. \end{cases}$$

Give $h = h_A$ and take the expectation of both sides of the equation (3), we have

$$F_{W_n}(A) - P_{\lambda_n}(A) = E\left[\lambda_n f(W_n + 1) - W_n f(W_n)\right]. \tag{4}$$

Thus, the problem of estimating Δ can be reduced to that of estimating the difference of the expectations

$$\left| E\lambda_n f(W_n + 1) - E W_n f(W_n) \right|.$$

According to Barbour et al. (see (Barbour, Holst & Janson, 1992), for $C_{w-1} = \{0, 1, ..., w-1\}$, the solution f_A of (3) is of the form

$$f_A(w) = \begin{cases} (w-1)! {\lambda_n}^{-w} e^{\lambda_n} \left[P_{\lambda_n}(h_{A \cap C_{w-1}}) - P_{\lambda_n}(h_A) P_{\lambda_n}(h_{C_{w-1}}) \right], & \text{if} \, w \geq 1, \\ 0, & \text{if} \, w = 0. \end{cases} \tag{5}$$

Before starting the main results in next section, we also need the following lemmas, which is directly obtained from (Barbour, Holst & Janson, 1992) and (Teerapabolarn & Wongkasem, 2007).

Lemma 1 *Let $Vf_A(w) = f_A(w+1) - f_A(w)$. Then, for $A \subseteq \mathbb{Z}_+$ and $k \in \mathbb{Z}_+ \setminus \{0\}$,*

$$\sup_{w \geq k} |Vf_A(w)| \leq \min\left\{ {\lambda_n}^{-1}\left(1 - e^{-\lambda_n}\right), \frac{1}{k} \right\}.$$

Lemma 2 *Let $w_0 \in \mathbb{Z}_+$ and $k \in \mathbb{Z}_+ \setminus \{0\}$, we have*

$$\sup_{w \geq k} \left| Vf_{C_{w_0}}(w) \right| \leq {\lambda_n}^{-1}\left(e^{\lambda_n} - 1\right) \min\left\{ \frac{1}{w_0 + 1}, \frac{1}{k} \right\}.$$

Lemma 3 *Let U_{λ_N} and U_λ denote a Poisson random variable with mean λ_N and λ, respectively. Then, for $A \subseteq \mathbb{Z}_+$, the total variation distance between the distributions of U_{λ_N} and U_λ satisfies the following inequality:*

$$d_{TV}(U_{\lambda_N}, U_\lambda) \leq \min\left\{ 1, \sqrt{\frac{2}{e\lambda}} \right\} E\left|\lambda_N - \lambda\right|. \tag{6}$$

3. Main Results

The following lemma is established for proving the main results.

Lemma 4 *Let $X_1, X_2, \ldots$ be a sequence of independent binomial distributed random variables. Setting $W_n = \sum_{i=1}^{n} X_i$ and $\lambda_n = E(W_n)$. Then,*

$$E\left[\lambda_n f\left(W_n+1\right)-W_n f\left(W_n\right)\right] = \sum_{i=1}^{n}\sum_{k\geq 1} kC_{r_i}^{k} p_i^{k+1}(1-p_i)^{r_i-k} E\left[f\left(W_i+k+1\right)-f\left(W_i+k\right)\right],$$

where f is a bounded real-valued function defined on $\mathbb{Z}_+$.

Proof. We have

$$E\left[\lambda_n f\left(W_n+1\right)-W_n f\left(W_n\right)\right] = \sum_{i=1}^{n} E\left[r_i p_i f\left(W_n+1\right)-X_i f\left(W_n\right)\right].$$

Setting $W_i = W_n - X_i$,

$$\begin{aligned}
&E\left[r_i p_i f\left(W_i+X_i+1\right)-X_i f\left(W_i+X_i\right)\right]\\
&= E\left[E\left[\left(r_i p_i f\left(W_i+X_i+1\right)-X_i f\left(W_i+X_i\right)\right)/X_i\right]\right]\\
&= E\left[r_i p_i f\left(W_i+1\right)\right] P\left(X_i=0\right)\\
&\quad + E\left[r_i p_i f\left(W_i+2\right)-f\left(W_i+1\right)\right] P\left(X_i=1\right)\\
&\quad + \sum_{k\geq 2} E\left[r_i p_i f\left(W_i+k+1\right)-k f\left(W_i+k\right)\right] P\left(X_i=k\right)\\
&= E\left[\left(r_i p_i P\left(X_i=0\right)-P\left(X_i=1\right)\right) f\left(W_i+1\right)\right]\\
&\quad + \sum_{k\geq 2} E\left[\left(r_i p_i P\left(X_i=k-1\right)-kP\left(X_i=k\right)\right) f\left(W_i+k\right)\right]\\
&= E\left[\left(r_i p_i(1-p_i)^{r_i}-r_i p_i(1-p_i)^{r_i-1}\right) f\left(W_i+1\right)\right]\\
&\quad + \sum_{k\geq 2} E\left[\left(r_i p_i C_{r_i}^{k-1} p_i^{k-1}(1-p_i)^{r_i-k+1}-kC_{r_i}^{k} p_i^{k}(1-p_i)^{r_i-k}\right) f\left(W_i+k\right)\right]\\
&= -E\left[r_i p_i^2(1-p_i)^{r_i-1} f\left(W_i+1\right)\right]\\
&\quad + \sum_{k\geq 2} E\left[\left(r_i p_i C_{r_i}^{k-1} p_i^{k-1}(1-p_i)^{r_i-k+1}-\left(r_i-k+1\right) C_{r_i}^{k-1} p_i^{k}(1-p_i)^{r_i-k}\right) f\left(W_i+k\right)\right]\\
&= -E\left[r_i p_i^2(1-p_i)^{r_i-1} f\left(W_i+1\right)\right]\\
&\quad + \sum_{k\geq 2} E\left[\left(\frac{r_i-k+1}{r_i} r_i p_i C_{r_i}^{k-1} p_i^{k-1}(1-p_i)^{r_i-k+1}-\left(r_i-k+1\right) C_{r_i}^{k-1} p_i^{k}(1-p_i)^{r_i-k}\right) f\left(W_i+k\right)\right]\\
&\quad - \sum_{k\geq 2} E\left[\left(\frac{r_i-k+1}{r_i}-1\right) r_i p_i C_{r_i}^{k-1} p_i^{k-1}(1-p_i)^{r_i-k+1} f\left(W_i+k\right)\right]\\
&= -E\left[r_i p_i^2(1-p_i)^{r_i-1} f\left(W_i+1\right)\right]\\
&\quad - \sum_{k\geq 2} E\left[\left(r_i-k+1\right) C_{r_i}^{k-1} p_i^{k+1}(1-p_i)^{r_i-k} f\left(W_i+k\right)\right]\\
&\quad - \sum_{k\geq 2} E\left[\left(\frac{r_i-k}{r_i}-1\right) r_i C_{r_i}^{k} p_i^{k+1}(1-p_i)^{r_i-k} f\left(W_i+k+1\right)\right]\\
&\quad + E\left[r_i p_i^2(1-p_i)^{r_i-1} f\left(W_i+2\right)\right]\\
&= r_i p_i^2(1-p_i)^{r_i-1} E\left[f\left(W_i+2\right)-f\left(W_i+1\right)\right]\\
&\quad + \sum_{k\geq 2} kC_{r_i}^{k} p_i^{k+1}(1-p_i)^{r_i-k} E\left[f\left(W_i+k+1\right)-f\left(W_i+k\right)\right]\\
&= \sum_{k\geq 1} kC_{r_i}^{k} p_i^{k+1}(1-p_i)^{r_i-k} E\left[f\left(W_i+k+1\right)-f\left(W_i+k\right)\right].
\end{aligned}$$

This finishes the proof. □

The following theorems present non-uniform and uniform bounds for the distance between the distribution functions of W_N and U_λ, which are the expected results.

3.1 A Uniform Bound on Poisson Approximation for Random Sums of Independent Binomial Random Variables

Theorem 1 *For $A \subseteq \mathbb{Z}_+$, we have*

$$\begin{aligned} d_{TV}(W_N, U_\lambda) \leq E\left(\sum_{i=1}^{N} \min\left\{\lambda_N^{-1}\left(1-e^{-\lambda_N}\right) r_i, \frac{1-(1-p_i)^{r_i}}{p_i}\right\} p_i^2\right) \\ + \min\left\{1, \sqrt{\frac{2}{\lambda e}}\right\} E\left|\lambda_N - \lambda\right|. \end{aligned} \tag{7}$$

Proof. Let $f = f_A$ be defined as in (5) and applying (4), we have

$$\left|P(W_n \in A) - \sum_{k\in A} \frac{\lambda_n^k e^{-\lambda_n}}{k!}\right| = \left|E[\lambda_n f(W_n+1) - W_n f(W_n)]\right|. \tag{8}$$

Taking account of Lemma 4 and Lemma 1, it follows that

$$\begin{aligned} &\left|E\left[r_i p_i f(W_n+1) - X_i f(W_n)\right]\right| \\ &\leq \sum_{k\geq 1} k C_{r_i}^k p_i^{k+1} (1-p_i)^{r_i-k} E\left|f(W_i+k+1) - f(W_i+k)\right| \\ &\leq \sum_{k\geq 1} k C_{r_i}^k p_i^{k+1} (1-p_i)^{r_i-k} \sup_{w\geq k} |Vf(w)| \\ &\leq \sum_{k\geq 1} k C_{r_i}^k p_i^{k+1} (1-p_i)^{r_i-k} \min\left\{\frac{1-e^{-\lambda_n}}{\lambda_n}, \frac{1}{k}\right\} \\ &= \min\left\{\frac{1-e^{-\lambda_n}}{\lambda_n} p_i \sum_{k\geq 1} k C_{r_i}^k p_i^k (1-p_i)^{r_i-k}, p_i \sum_{k\geq 1} C_{r_i}^k p_i^k (1-p_i)^{r_i-k}\right\} \\ &= \min\left\{\frac{1-e^{-\lambda_n}}{\lambda_n} p_i \sum_{k\geq 1} k P(X_i=k), p_i\left(\sum_{k\geq 0} P(X_i=k) - (1-p_i)^{r_i}\right)\right\} \\ &= \min\left\{\frac{1-e^{-\lambda_n}}{\lambda_n} p_i E(X_i), p_i\left(1-(1-p_i)^{r_i}\right)\right\}. \end{aligned}$$

Thus,

$$\left|E\left[r_i p_i f(W_n+1) - X_i f(W_n)\right]\right| \leq \min\left\{\lambda_n^{-1}\left(1-e^{-\lambda_n}\right) r_i, \frac{1-(1-p_i)^{r_i}}{p_i}\right\} p_i^2. \tag{9}$$

Combining (8) with (9), gives

$$d_{TV}(W_n, U_{\lambda_n}) \leq \sum_{i=1}^{n} \min\left\{\lambda_n^{-1}\left(1-e^{-\lambda_n}\right) r_i, \frac{1-(1-p_i)^{r_i}}{p_i}\right\} p_i^2. \tag{10}$$

From Lemma 3 and (10), it follows the fact that

$$\begin{aligned} d_{TV}(W_N, U_\lambda) &= \sum_{n=1}^{\infty} P(N=n)\, d_{TV}(W_n, U_\lambda) \\ &\leq \sum_{n=1}^{\infty} P(N=n)\left[d_{TV}(W_n, U_{\lambda_n}) + d_{TV}(U_{\lambda_n}, U_\lambda)\right] \\ &= \sum_{n=1}^{\infty} P(N=n)\, d_{TV}(W_n, U_{\lambda_n}) + d_{TV}(U_{\lambda_N}, U_\lambda) \\ &\leq \sum_{n=1}^{\infty} P(N=n) \sum_{i=1}^{n} \min\left\{\lambda_n^{-1}\left(1-e^{-\lambda_n}\right) r_i, \frac{1-(1-p_i)^{r_i}}{p_i}\right\} p_i^2 \\ &\quad + \min\left\{1, \sqrt{\frac{2}{\lambda e}}\right\} E\left|\lambda_N - \lambda\right| \end{aligned}$$

$$\leq E\left(\sum_{i=1}^{N}\min\left\{\lambda_N^{-1}\left(1-e^{-\lambda_N}\right)r_i,\frac{1-(1-p_i)^{r_i}}{p_i}\right\}p_i^2\right)$$
$$+\min\left\{1,\sqrt{\frac{2}{\lambda e}}\right\}E\left|\lambda_N-\lambda\right|.$$

This finishes the proof. □

Remark 1 For $r_1 = r_2 = ... = r_n = 1$, we have a uniform bound on Poisson approximation for the random sums of independent Bernoulli random variables:

$$d_{TV}\left(V_N,U_{\lambda^*}\right)\leq E\left(\lambda^{*}{}_N^{-1}\left(1-e^{-\lambda_N^*}\right)\sum_{i=1}^{N}p_i^2\right)+\min\left\{1,\sqrt{\frac{2}{\lambda^* e}}\right\}E\left|\lambda_N^*-\lambda^*\right|. \tag{11}$$

Remark 2 Let us consider:

$$\min\left\{1,\sqrt{\frac{2}{\lambda^* e}}\right\}\leq 1$$

and

$$\min\left\{\frac{1-e^{\lambda_N}}{\lambda_N}r_i,\frac{1-(1-p_i)^{r_i}}{p_i}\right\}p_i^2\leq\frac{1-e^{\lambda_N}}{\lambda_N}r_ip_i^2.$$

Thus, the bounds in (7) and (11) are sharper than the bounds in (2) and (1), respectively.

Corollary 1 *For $N = n \in \mathbb{Z}_+$ is fixed, then $\lambda = \lambda_n = \sum_{i=1}^{n} r_i p_i$ and*

$$d_{TV}\left(W_n,U_{\lambda_n}\right)\leq\sum_{i=1}^{n}\min\left\{\lambda_n^{-1}\left(1-e^{-\lambda_n}\right)r_i,\frac{1-(1-p_i)^{r_i}}{p_i}\right\}p_i^2. \tag{12}$$

Remark 3 The result (12) is a uniform bound on Poisson approximation for sums of independent binomial random variables. This bound is sharper than those reported in (Teerapabolarn, 2014a).

3.2 A Non-uniform Bound on Poisson Approximation for Random Sums of Independent Binomial Random Variables

Theorem 2 *For $w_0 \in \mathbb{Z}_+$, we have*

$$\begin{aligned}\left|P(W_N\leq w_0)-P\left(U_\lambda\leq w_0\right)\right|&\leq\min\left\{\frac{2\lambda}{w_0+1},\min\left\{1,\sqrt{\frac{2}{e\lambda}}\right\}E\left|\lambda_N-\lambda\right|\right\}\\&+E\left(\sum_{i=1}^{N}{\lambda_N}^{-1}\left(1-e^{-\lambda_N}\right)\min\left\{\frac{e^{\lambda_N}r_i}{(w_0+1)},\frac{\left(1-(1-p_i)^{r_i}\right)e^{\lambda_N}}{p_i}\right\}p_i^2\right).\end{aligned} \tag{13}$$

Proof. For $C_w = \{0, ..., w\}$ and $w_0 \in \mathbb{Z}_+$, let $h_{w_0} : \mathbb{Z}_+ \to \mathbb{R}$ such that

$$h_{C_{w_0}}(w)=\begin{cases}1 & \text{if } w\leq w_0,\\ 0 & \text{if } w>w_0.\end{cases}$$

According to Barbour et al. (see (Barbour, Holst & Janson, 1992) on p.7), the solution $f_{C_{w_0}}(w)$ of (3) is expressed in the form of

$$f_{C_{w_0}}(w)=\begin{cases}(w-1)!{\lambda_n}^{-w}e^{\lambda_n}\left[P_{\lambda_n}\left(h_{C_{w_0}}\right)P_{\lambda_n}\left(1-h_{C_{w-1}}\right)\right] & ,\text{if} \quad w_0<w,\\ (w-1)!{\lambda_n}^{-w}e^{\lambda_n}\left[P_{\lambda_n}\left(h_{C_{w-1}}\right)P_{\lambda_n}\left(1-h_{C_{w_0}}\right)\right] & ,\text{if} \quad w_0\geq w,\\ 0 & ,\text{if} \quad w=0.\end{cases}$$

Given $f = f_{C_{w_0}}$ and $h = h_{C_{w_0}}$, the Stein's equation

$$h_{C_{w_0}}(w)-\sum_{k\leq w_0}e^{-\lambda_n}\frac{\lambda_n^k}{k!}=\lambda_n f(w+1)-wf(w).$$

Taking expectations of both sides, and applying Lemma 2 and Lemma 4, we have

$$\begin{aligned}
&\left|P(W_n \le w_0) - P\left(U_{\lambda_n} \le w_0\right)\right| \\
&\le \sum_{i=1}^{n}\left(\sum_{k\ge 1} kC_r^k p_i^{k+1}(1-p_i)^{r_i-k} E\left|f\left(W_i+k+1\right) - f\left(W_i+k\right)\right|\right) \\
&\le \sum_{i=1}^{n}\left(\sum_{k\ge 1} kC_r^k p_i^{k+1}(1-p_i)^{r_i-k} \lambda_n^{-1}\left(e^{\lambda_n}-1\right)\min\left\{\frac{1}{w_0+1}, \frac{1}{k}\right\}\right) \\
&= \sum_{i=1}^{n} \lambda_n^{-1}\left(e^{\lambda_n}-1\right)\min\left\{\frac{p_i \sum\limits_{k\ge 1} kP\left(X_i=k\right)}{w_0+1}, p_i\sum_{k\ge 1} P\left(X_i=k\right)\right\} \\
&= \lambda_n^{-1}\left(e^{\lambda_n}-1\right)\sum_{i=1}^{n}\min\left\{\frac{r_i}{w_0+1}, \frac{1-(1-p_i)^{r_i}}{p_i}\right\} p_i^2.
\end{aligned}$$

Thus,

$$\left|P(W_n \le w_0) - P\left(U_{\lambda_n} \le w_0\right)\right| \le \lambda_n^{-1}\left(e^{\lambda_n}-1\right)\sum_{i=1}^{n}\min\left\{\frac{r_i}{w_0+1}, \frac{1-(1-p_i)^{r_i}}{p_i}\right\} p_i^2. \tag{14}$$

In addition, by using Lemma 3, Teerapabolarn showed that (see (Teerapabolarn, 2013) for more details):

$$\left|P\left(U_{\lambda_N} \le w_0\right) - P\left(U_{\lambda} \le w_0\right)\right| \le \min\left\{\frac{2\lambda}{w_0+1}, \min\left\{1, \sqrt{\frac{2}{e\lambda}}\right\} E\left|\lambda_N - \lambda\right|\right\}. \tag{15}$$

Combining (14) and (15) gives

$$\begin{aligned}
&\left|P\left(W_N \le w_0\right) - P\left(U_{\lambda} \le w_0\right)\right| \\
&\le \sum_{n=0}^{\infty} P\left(N=n\right)\left|P\left(W_n \le w_0\right) - P\left(U_{\lambda} \le w_0\right)\right| \\
&\le \sum_{n=0}^{\infty} P\left(N=n\right)\left|P\left(W_n \le w_0\right) - P\left(U_{\lambda_n} \le w_0\right)\right| \\
&\quad + \left|P\left(U_{\lambda_N} \le w_0\right) - P\left(U_{\lambda} \le w_0\right)\right| \\
&\le \sum_{n=0}^{\infty} P\left(N=n\right)\frac{1-e^{-\lambda_n}}{\lambda_n}\sum_{i=1}^{n}\min\left\{\frac{r_i e^{\lambda_n}}{w_0+1}, \frac{\left(1-(1-p_i)^{r_i}\right)e^{\lambda_n}}{p_i}\right\} p_i^2 \\
&\quad + \min\left\{\frac{2\lambda}{w_0+1}, \min\left\{1, \sqrt{\frac{2}{e\lambda}}\right\} E\left|\lambda_N - \lambda\right|\right\} \\
&\le E\left(\frac{1-e^{-\lambda_N}}{\lambda_N}\sum_{i=1}^{N}\min\left\{\frac{r_i e^{\lambda_N}}{w_0+1}, \frac{\left(1-(1-p_i)^{r_i}\right)e^{\lambda_N}}{p_i}\right\} p_i^2\right) \\
&\quad + \min\left\{\frac{2\lambda}{w_0+1}, \min\left\{1, \sqrt{\frac{2}{e\lambda}}\right\} E\left|\lambda_N - \lambda\right|\right\}.
\end{aligned}$$

This finishes the proof. □

Remark 4 For $r_1 = r_2 = ... = r_n = 1$, we have a non-uniform bound on Poisson approximation for the random sums of independent Bernoulli random variables:

$$\begin{aligned}
\left|P(V_N \le w_0) - P\left(U_{\lambda^*} \le w_0\right)\right| \le \min&\left\{\frac{2\lambda^*}{w_0+1}, \min\left\{1, \sqrt{\frac{2}{e\lambda^*}}\right\} E\left|\lambda_N^* - \lambda^*\right|\right\} \\
&+ E\left(\frac{\left(e^{\lambda_N^*}-1\right)}{\left(w_0+1\right)\lambda_N^*}\sum_{i=1}^{N} p_i^2\right).
\end{aligned} \tag{16}$$

Corollary 2 *For $N = n \in \mathbb{Z}_+$ is fixed, then $\lambda = \lambda_n = \sum_{i=1}^{n} r_i p_i$ and*

$$\left|P(W_n \leq w_0) - P\left(U_{\lambda_n} \leq w_0\right)\right| \leq \lambda_n^{-1}\left(e^{\lambda_n} - 1\right)\sum_{i=1}^{n} \min\left\{\frac{r_i}{w_0+1}, \frac{1-(1-p_i)^{r_i}}{p_i}\right\} p_i^2. \tag{17}$$

Remark 5 The result (17) is a non-uniform bound on Poisson approximation for sums of independent binomial random variables.

4. Conclusions

We conclude this paper with the following comments. Bounds for the distance between the distribution function of random sums of independent binomial random variables and an appropriate Poisson distribution function were obtained. The received results in this paper are sharper than those reported in (Teerapabolarn, 2014a), (Teerapabolarn, 2014b), and (Yannaros, 1991). Moreover, non-uniform bounds on Poisson approximation for sums (and random sums) of independent binomial random variables are given. The results will be more interesting and valuable if we discuss Poisson approximation for random sums of dependent binomial random variables. We shall continue studying this matter in our future research.

References

Barbour, A. D., Holst, L., & Janson, S. (1992). *Poisson Approximation*. Clarendon Press-Oxford.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probab., 3*(3), 534–545.

Hung, L. T., & Giang, T. L. (2014). On bounds in Poisson approximation for integer-valued independent random variables. *Journal of Inequalities and Applications.* https://doi.org/10.1186/1029-242X-2014-291

Hung, L. T., & Giang, T. L. (2016a). *On bounds in Poisson approximation for distributions of independent negative-binomial distributed random variables*. SpringerPlus. https://doi.org/10.1186/s40064-016-1710-y

Hung, L. T., & Giang, T. L. (2016b). On the bounds in Poisson approximation for independent geometric distributed random variables. *Bulletin of the Irannian Mathematical Society, 42*(5), 1087–1096.

Kongudomthrap, S., & Chaidee, N. (2012). Bounds in Poisson approximation of random sums of Bernoulli random variables. *Journal of Mathematics Research*, (4), 29–35.

Teerapabolarn K., & Wongkasem P. (2007). Poisson approximation for independent geometric random variables. *Int. Math. Forum, 2*, 3211–3218.

Teerapabolarn, K. (2013). Improved bounds on Poisson approximation for independent binomial random summands. *International Journal of Pure and Applied Mathematics, 89*(1), 29–33.

Teerapabolarn, K. (2014a). Poisson approximation for independent binomial random variables. *International Journal of Pure and Applied Mathematics, 93*(6), 775–777.

Teerapabolarn, K. (2014b). Poisson approximation for random sums of independent binomial random variables. *Applied Mathematical Sciences, 8*(173), 8643–8646.

Vellaisamy, P., & Upadhye, S. (2009). Compound negative binomial approximations for sums of random variables. *Probab. Math. Statist.*, (29), 205 - 226.

Yannaros, N. (1991). Poisson approximation for random sums of Bernoulli random variables. *Statistics & Probability Letters*, (11), 161–165.

Copyrights

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

On Optimal Allocation of Treatment/Condition Variance in Principal Component Analysis

André Beauducel[1] & Norbert Hilger[1]

[1] Department of Psychology, University of Bonn, Germany

Correspondence: André Beauducel, Department of Psychology, University of Bonn, Kaiser-Karl-Ring, 9, 53111 Bonn, Germany. E-mail: beauducel@uni-bonn.de

Received: April 23, 2018 Accepted: May 8, 2018 Online Published: June 19, 2018

doi:10.5539/ijsp.v7n4p50 URL: https://doi.org/10.5539/ijsp.v7n4p50

Abstract

The allocation of a (treatment) condition-effect on the wrong principal component (misallocation of variance) in principal component analysis (PCA) has been addressed in research on event-related potentials of the electroencephalogram. However, the correct allocation of condition-effects on PCA components might be relevant in several domains of research. The present paper investigates whether different loading patterns at each condition-level are a basis for an optimal allocation of between-condition variance on principal components. It turns out that a similar loading shape at each condition-level is a necessary condition for an optimal allocation of between-condition variance, whereas a similar loading magnitude is not necessary.

Keywords: Principal component analysis, misallocation of variance, within- and between condition effects

1. Introduction

1.1 Condition Effects in Principal Component Analysis

Principal component analysis (PCA) has regularly been performed for the analysis of event-related potentials of the electroencephalogram (Dien, Khoe & Mangun, 2007; Dien, 2010; Kayser & Tenke, 2003, 2005). In the context of event-related potentials, PCA is often performed for observed variables representing *k* levels of at least one (experimental) condition factor, so that the components represent a mixture of the between- and within-condition variance. However, (experimental) condition factors occur in several areas of research and PCA is performed in several areas of research and has been adapted to several different methodological contexts (Jolliffe & Cadima, 2016). It is therefore interesting to know how experimental condition effects are optimally allocated on principal components.

1.2 Misallocation of Between-condition Variance

Since Wood and McCarthy (1984) it has been regarded as an optimum when a single PCA component combines the complete between-condition variance of a single condition factor with some within-condition variance. Although the allocation of the variance of a single condition factor on a single principal component combining within- and between-condition variance was described as an optimum in research on event-related potentials, this form of variance allocation might also be useful in other contexts of research as it allows for a parsimonious data description. Wood and McCarthy (1984) used the term ‘misallocation of variance’ in order to denote that the between-condition variance is allocated on more than one principal component. Misallocation of variance has been investigated in simulation studies on methods of PCA component rotation (e.g., Scharf & Nestler, in press; Dien, 2010; Beauducel & Debener, 2003; Wood & McCarthy, 1984) and new methods of component rotation have been proposed that may reduce misallocation of variance (Beauducel, 2018; Beauducel & Leue, 2015). Moreover, Scharf and Nestler (in press) have demonstrated that the covariation of several condition factors may induce misallocation of variance.

It has also been proposed to perform a separate PCA for each group representing a level of a single condition factor because the loading shapes in each condition can be different (Barry, De Blasio, Fogarty & Karamacoska, 2016). Although it might be reasonable to identify condition-specific loading patterns by means of separate PCAs at each level of a condition factor, the effect of this form of analysis on misallocation of variance remains unknown.

1.3 Aims of the Present Paper

The present paper therefore investigates the effects of separate PCAs at each level of a single condition factor on the allocation of between-condition variance on PCA components. First, some definitions for separate PCAs at each level of a single condition factor and for a PCA of the between-condition variance of the condition factor are presented. Second, it is shown that misallocation of condition variance as it has been demonstrated and discussed since Wood and McCarthy (1984) follows necessarily from rotation of components that perfectly represent a single condition effect. Third, it is shown that different condition-specific loading shapes do not allow for an unambiguous allocation of between-condition variance on a single component representing within- and between-condition variance. Finally, it is shown that different condition-specific loading patterns are compatible with an unambiguous allocation of between-condition variance on a single component, when the between-condition differences of the loadings on each component can be accounted for by a scalar.

2. Definitions: PCA for within- and Between-condition Variance

Consider that p random variables have been observed in k levels of a condition factor, so that

$$\mathbf{x} = \left[\mathbf{x}_i \quad \cdots \quad \mathbf{x}_k\right], \ \textit{for } i = 1, 2, \ldots, k. \tag{1}$$

Although the expectation of $\mathbf{x}$ is zero $(\mathrm{E}[\mathbf{x}] = 0)$, the conditions imply $\mathrm{E}[\mathbf{x}_i] \neq 0$. However, when a within- condition PCA is performed separately for the correlations or covariances at each level of the condition factor, $\mathbf{x}_i^v = \mathbf{x}_i - \mathrm{E}[\mathbf{x}_i]$, the mean centered part of $\mathbf{x}_i$, is analyzed, since $\mathrm{Cov}[\mathbf{x}_i, \mathbf{x}_i] = \mathrm{Cov}[\mathbf{x}_i^v, \mathbf{x}_i^v] = \mathbf{\Sigma}_i$, so that

$$\mathbf{x}_i^v = \mathbf{A}_i^v \mathbf{c}_i^v, \ \textit{for } i = 1, 2, \ldots k, \tag{2}$$

superscript "v" denotes the within-condition variance and where $\mathbf{A}_i^v$ is a $p \times p$ matrix of component loadings and $\mathbf{A}_i^{v'}\mathbf{A}_i^v$ contains the eigenvalues in decreasing order. The components $\mathbf{c}_i^v$ are assumed to have an expectation zero, $\mathrm{E}[\mathbf{c}_i^v] = 0$. PCA initially yields orthogonal components $(\mathrm{E}[\mathbf{c}_i^v \mathbf{c}_i^{v'}] = \mathbf{I})$, so that each covariance matrix of observed variables can be decomposed into

$$\sum\nolimits_i^v = \mathbf{A}_i^v \mathbf{A}_i^{v'}, \ \textit{for } i = 1, 2, \ldots k, \tag{3}$$

Typically, components $\mathbf{c}_i$ are divided into a subset of q wanted components $\mathbf{w}_i$ and $p - q$ unwanted components $\mathbf{u}_i$ ($\mathbf{c}_i = [\mathbf{w}_i, \mathbf{u}_i]$, $\mathbf{A}_i = [\mathbf{M}_i, \mathbf{N}_i]$). Orthogonal and oblique rotations of $\mathbf{M}_i$ and $\mathbf{w}_i$ have been proposed, so that non-zero component inter-correlations are possible $(\mathrm{E}[\mathbf{c}_i \mathbf{c}_i'] = \mathbf{Q}_i)$. The covariances of observed variables are then decomposed by

$$\mathbf{\Sigma}_i = \mathbf{M}_i \mathbf{Q}_i \mathbf{M}_i' + \mathbf{N}_i \mathbf{N}_i', \ \textit{for } i = 1, 2, \ldots k. \tag{4}$$

It is possible to write the complete data comprising condition variance and within-group variance as

$$\begin{aligned} \mathbf{x} &= \left[\mathbf{x}_i^v + diag(\mathrm{E}[\mathbf{x}_i])\mathbf{1}_i \ , \quad \cdots \quad , \mathbf{x}_k^v + diag(\mathrm{E}[\mathbf{x}_k])\mathbf{1}_k\right] \\ &= \left[\mathbf{x}_i^v \quad \cdots \quad \mathbf{x}_k^v\right] + \left[diag(\mathrm{E}[\mathbf{x}_i])\mathbf{1}_i \ , \quad \cdots \quad , diag(\mathrm{E}[\mathbf{x}_k])\mathbf{1}_k\right] \\ &= \left[\mathbf{x}_i^v \quad \cdots \quad \mathbf{x}_k^v\right] + \left[\mathbf{x}_i^b \quad \cdots \quad \mathbf{x}_k^b\right], \end{aligned} \tag{5}$$

where $\mathbf{1}_i$ has the dimensions of $\mathbf{x}_i$ and $\mathbf{1}_k$ has the dimensions of $\mathbf{x}_k$. The related within- and between-conditions PCAs yield

$$\mathbf{x} = \left[\mathbf{A}_i^v \mathbf{c}_i^v, \quad \cdots \quad , \mathbf{A}_k^v \mathbf{c}_k^v\right] + \mathbf{A}^b \mathbf{c}^b. \tag{6}$$

Usually q_v wanted within-condition components $\mathbf{w}_i^v$ are separated from $p - q_v$ unwanted within-condition components $\mathbf{u}_i^v$ and q_b wanted between-condition components $\mathbf{w}^b$ from $p - q_b$ unwanted between-condition components $\mathbf{u}^b$. This yields

$$\mathbf{x} = \left[\mathbf{M}_i^v \mathbf{w}_i^v + \mathbf{N}_i^v \mathbf{u}_i^v, \quad \cdots \quad , \mathbf{M}_k^v \mathbf{w}_k^v + \mathbf{N}_k^v \mathbf{u}_k^v\right] + \mathbf{M}^b \mathbf{w}^b + \mathbf{N}^b \mathbf{u}^b, \tag{7}$$

and

$$\mathbf{x}_w = \left[\mathbf{M}_i^v \mathbf{w}_i^v, \quad \cdots \quad , \mathbf{M}_k^v \mathbf{w}_k^v\right] + \mathbf{M}^b \mathbf{w}^b, \tag{8}$$

for the wanted components.

Typically, the wanted components are rotated in order to improve the interpretation (Dien, 2010; Kayser & Tenke, 2003). If there is an additional condition factor, there can be additional groupings of PCAs for each level of the condition factor and an additional PCA across the levels of the condition factor. If the sample size is sufficiently large, it is also possible to perform a PCA for each of the combinations of condition levels and across all combinations of conditions of the two condition factors.

3. Misallocation of Variance

3.1 Misallocation of Variance and Component Rotation

When there are only a few condition factors the number of wanted within-condition components is probably larger than the number of wanted between-condition components. For example, when there is only one condition factor with two levels, PCA of the between-condition variance without subsequent component rotation will result in only one between-condition component. When $q^v > q^b = 1$ it is possible to write Equation 8 as

$$\mathbf{x}_w = \left[\begin{bmatrix} \mathbf{m}^v_{ji}, & \cdots & ,\mathbf{m}^v_{qi} \end{bmatrix} \begin{bmatrix} \mathbf{w}^v_{ji} \\ \vdots \\ \mathbf{w}^v_{qi} \end{bmatrix}, \quad \cdots \quad , \begin{bmatrix} \mathbf{m}^v_{jk}, & \cdots & ,\mathbf{m}^v_{qk} \end{bmatrix} \begin{bmatrix} \mathbf{w}^v_{jk} \\ \vdots \\ \mathbf{w}^v_{qk} \end{bmatrix} \right] + \mathbf{m}^b \mathbf{w}^b, \tag{9}$$

where j denotes the number of the respective within-condition component. For $q^b = 1$ and $\mathbf{m}^v_{1i} = \mathbf{m}^b, \quad \cdots \quad , \mathbf{m}^v_{1k} = \mathbf{m}^b$ Equation 9 can be written as

$$\mathbf{x}_w = \mathbf{M} \left[\begin{bmatrix} \mathbf{w}_{ji} \\ \vdots \\ \mathbf{w}_{qi} \end{bmatrix}, \quad \cdots \quad , \begin{bmatrix} \mathbf{w}_{jk} \\ \vdots \\ \mathbf{w}_{qk} \end{bmatrix} \right], \tag{10}$$

$$with\ \mathbf{M} = \begin{bmatrix} \mathbf{m}^b, \mathbf{m}^v_{2i}, & \cdots & ,\mathbf{m}^v_{qi} \end{bmatrix}\ and\ \begin{bmatrix} \mathbf{w}_{1i}, \cdots, \mathbf{w}_{1k} \end{bmatrix} = \begin{bmatrix} \mathbf{w}^v_{1i} + \mathbf{w}^b_{1i}, \cdots, \mathbf{w}^v_{1k} + \mathbf{w}^b_{1k} \end{bmatrix},$$

where $\mathbf{w}^v_{1i}, \cdots, \mathbf{w}^v_{1k}$ denotes the scores on the first wanted component (j=1) at each level i of the condition factor, and $\mathbf{w}^b_{1i}, \cdots, \mathbf{w}^b_{1k}$ denotes the expectancy of the first wanted component on each level i of the condition factor, which corresponds to the expectancy of the observed scores on condition level i, with $[\mathbf{w}^b_{1i}, \cdots, \mathbf{w}^b_{1k}] = [\mathrm{E}(\mathbf{w}_{1i}), \cdots, \mathrm{E}(\mathbf{w}_{1k})] = [\mathrm{E}(\mathbf{x}_i), \cdots, \mathrm{E}(\mathbf{x}_k)]$.

Equation 10 describes what is typically regarded as an optimal allocation of variance, namely, that a condition effect occurs on a single component that combines within- and between-condition variance. The simulation studies on this issue were based on a single condition effect that was introduced exclusively on a single component when the data were generated (Wood & McCarthy, 1984; Dien, 2010; Beauducel & Debener, 2003; Beauducel & Leue, 2015) and that occurred on more than one component after PCA followed by component rotation.

Component rotation means that the **M** is rotated by means of postmultiplication by a $q^v \times q^v$ transformation matrix **T** (Harman, 1976) and that the component scores are counter-rotated by means of premultiplication with $\mathbf{T}^{-1}$, so that

$$\mathbf{x}_w = \mathbf{MT} \left[\mathbf{T}^{-1} \begin{bmatrix} \mathbf{w}_{ji} \\ \vdots \\ \mathbf{w}_{qi} \end{bmatrix}, \quad \cdots \quad , \mathbf{T}^{-1} \begin{bmatrix} \mathbf{w}_{jk} \\ \vdots \\ \mathbf{w}_{qk} \end{bmatrix} \right], \ with \left[\mathbf{w}_{1i}, \cdots, \mathbf{w}_{1k} \right] = \left[\mathbf{w}^v_{1i} + \mathbf{w}^b_{1i}, \cdots, \mathbf{w}^v_{1k} + \mathbf{w}^b_{1k} \right]. \tag{11}$$

For a single condition i the rotation of the infinite matrices containing the population of individual component scores l can be written as

$$\mathbf{T}^{-1} \begin{bmatrix} \mathbf{w}_{ji} \\ \vdots \\ \mathbf{w}_{qi} \end{bmatrix} = \mathbf{T}^* \begin{bmatrix} \mathbf{w}_{jil}, \cdots \\ \vdots \quad \ddots \\ \mathbf{w}_{qil}, \cdots \end{bmatrix} = \begin{bmatrix} \left(\mathbf{t}^* \mathbf{w} \right)_{jil}, \cdots \\ \vdots \quad \ddots \\ \left(\mathbf{t}^* \mathbf{w} \right)_{qil}, \cdots \end{bmatrix}, \tag{12}$$

$$for\ l = 1, ..., \infty,\ with \left[\mathbf{w}_{1il}, \quad \cdots \right] = \left[\mathbf{w}^v_{1il} + \mathrm{E}\left(\mathbf{w}_{1i} \right), \quad \cdots \right]\ and\ \mathbf{T}^{-1} = \mathbf{T}^*.$$

Theorem 1 describes that a non-zero expectation that is initially only on the first component leads to a non-zero expectation on others than the first component after component rotation.

Theorem 1. $If\ \mathrm{E}\left(\mathbf{w}_{ji} \right) = \begin{cases} \mathrm{E}\left(\mathbf{w}_{ji} \right) \neq 0\ for\ j = 1 \\ \mathrm{E}\left(\mathbf{w}_{ji} \right) = 0\ for\ j = 2, ..., q \end{cases}$ $and\ \mathbf{t}^*_{jh} \neq 0, for\ j = 1, ..., q, h = 1, ..., q,$

$then\ \mathrm{E}\left(\mathbf{t}^* \mathbf{w} \right)_{ji} \neq 0,\ for\ j > 1, ..., q.$

Proof. A single element for condition *i* of the matrix resulting from Equation 12 is given by

$$\left(\mathbf{t}^{*}\mathbf{w}\right)_{jil} = \sum_{h=1}^{q} \mathbf{t}^{*}_{jh}\mathbf{w}_{hil},\ with\ \mathbf{w}_{1il} = \mathbf{w}^{v}_{1il} + \mathrm{E}\left(\mathbf{w}_{1i}\right). \tag{13}$$

Equation 13 can be written as

$$\left(\mathbf{t}^{*}\mathbf{w}\right)_{jil} = \mathbf{t}^{*}_{j1}\left(\mathbf{w}^{v}_{1il} + \mathrm{E}\left(\mathbf{w}_{1i}\right)\right) + \ldots + \mathbf{t}^{*}_{jq}\mathbf{w}_{qil}. \tag{14}$$

Equation 14 implies that the expectation for the population of scores even for $j > 1$ is $\mathrm{E}\left(\mathbf{t}^{*}\mathbf{w}\right)_{ji} = \mathbf{t}^{*}_{j1}\mathrm{E}\left(\mathbf{w}_{1i}\right)$.

This completes the proof. □

Theorem 1 implies that a condition effect that occurs only on the first component before rotation, also occurs on other components after rotation. Thus, Theorem 1 shows that misallocation of variance as it has typically been investigated in simulation studies since Wood and McCarthy (1984) is a necessary consequence of any rotation of an initial set of components combining unambiguously within- and between-condition effects. Therefore, the attempts to reduce misallocation of variance are attempts to recover the initial combination of within- and between-condition components (Dien, 2010; Beauducel & Leue, 2015; Beauducel, 2018) so that the matrix **T**, transforming the original components to the given components becomes **I**. This implies $\mathbf{T}^{*} = \mathbf{I}$ and $\mathbf{t}^{*}_{jh} = 0, for\ j \neq h$ so that Theorem 1 does not hold. Eliminating variance misallocation by means of component rotation precludes that there exists a PCA solution for the data at hand where each between-condition effect can be allocated on a separate single component. This is, however, not necessarily the case for any data set.

3.2 Misallocation of Variance in Combined within- and Between-Condition Components

Theorem 1 describes misallocation of variance as it can occur when PCA is performed for the total sample, i.e., across the levels of a between-condition factor. When separate within-condition components $\mathbf{c}^{v}_{i},\ldots,\mathbf{c}^{v}_{k}$ are computed, the within-condition components $\mathbf{c}^{v}_{i},\ldots,\mathbf{c}^{v}_{k}$ are completely unrelated to $\mathbf{c}^{b}$ so that within- and between-condition variance is completely disentangled. This yields the question under which constraints within- and between-condition components can be combined into a single component representing within- and between-condition variance unambiguously. Theorem 2 describes a constraint for the component loadings that implies $\mathbf{c} = [\mathbf{c}^{v}_{i} + \mathbf{c}^{b}_{i},\ \cdots\ , \mathbf{c}^{v}_{k} + \mathbf{c}^{b}_{k}]$, i.e., that each component in **c** can be decomposed into a separate within- and between-condition component. This implies that no misallocation of variance occurs because each between-condition component is uniquely combined with another within-condition component.

Theorem 2. *If* $\mathbf{A}^{v}_{i} = \mathbf{A}^{b}, \cdots, \mathbf{A}^{v}_{k} = \mathbf{A}^{b}$, *then* $\mathbf{c} = \left[\mathbf{c}^{v}_{i},\ \cdots\ , \mathbf{c}^{v}_{k}\right] + \left[\mathbf{c}^{b}_{i},\ \cdots\ , \mathbf{c}^{b}_{k}\right]$.

Proof. Since $\mathbf{c}^{b} = [\mathbf{c}^{b}_{i}, \ldots, \mathbf{c}^{b}_{k}]$ Equation 6 can be written as

$$\mathbf{x} = \left[\mathbf{A}^{v}_{i}\mathbf{c}^{v}_{i},\ \cdots\ , \mathbf{A}^{v}_{k}\mathbf{c}^{v}_{k}\right] + \mathbf{A}^{b}\left[\mathbf{c}^{b}_{i},\ \cdots\ , \mathbf{c}^{b}_{k}\right], \tag{15}$$

inserting $\mathbf{A}^{b}$ for $\mathbf{A}^{v}_{i}, \cdots, \mathbf{A}^{v}_{k}$ into Equation 15 yields

$$\begin{aligned} \mathbf{x} &= \mathbf{A}^{b}\left[\mathbf{c}^{v}_{i},\ \cdots\ , \mathbf{c}^{v}_{k}\right] + \mathbf{A}^{b}\left[\mathbf{c}^{b}_{i},\ \cdots\ , \mathbf{c}^{b}_{k}\right] \\ &= \mathbf{A}^{b}\left[\mathbf{c}^{v}_{i} + \mathbf{c}^{b}_{i},\ \cdots\ , \mathbf{c}^{v}_{k} + \mathbf{c}^{b}_{k}\right] = \mathbf{A}^{b}\mathbf{c}. \end{aligned} \tag{16}$$

This completes the proof. □

Thus, when the within-condition loading matrices at each condition level are identical to the between-condition loading matrix, this implies a component model where all components combine their respective within- and between-condition variance. Theorem 2 implies that no misallocation of variance occurs when each condition-specific loading pattern is identical to the between-condition loading pattern. When Theorem 2 holds, it would be possible to find a solution without variance misallocation by means of component rotation. Writing loading vectors $\mathbf{A}^{v}_{i} = [\mathbf{a}^{v}_{si}, \cdots, \mathbf{a}^{v}_{pi}], \mathbf{A}^{b} = [\mathbf{a}^{b}_{s}, \cdots, \mathbf{a}^{b}_{p}]$ and component score

vectors $\mathbf{c}^{v}_{i} = \begin{bmatrix} \mathbf{c}^{v}_{si} \\ \vdots \\ \mathbf{c}^{v}_{pi} \end{bmatrix}, \mathbf{c}^{b}_{i} = \begin{bmatrix} \mathbf{c}^{b}_{si} \\ \vdots \\ \mathbf{c}^{b}_{pi} \end{bmatrix}$ for the *s* to *p* components in Equation 16 yields

$$\mathbf{x} = \left[\begin{bmatrix} & & \\ [\mathbf{a}_{si}^{v}, \cdots, \mathbf{a}_{pi}^{v}] & & \end{bmatrix} \begin{bmatrix} \mathbf{c}_{si}^{v} \\ \vdots \\ \mathbf{c}_{pi}^{v} \end{bmatrix}, \quad \cdots \quad , [\mathbf{a}_{sk}^{v}, \cdots, \mathbf{a}_{pk}^{v}] \begin{bmatrix} \mathbf{c}_{sk}^{v} \\ \vdots \\ \mathbf{c}_{pk}^{v} \end{bmatrix} \right] + [\mathbf{a}_{s}^{b}, \cdots, \mathbf{a}_{p}^{b}] \left[\begin{bmatrix} \mathbf{c}_{si}^{b} \\ \vdots \\ \mathbf{c}_{pi}^{b} \end{bmatrix}, \quad \cdots \quad , \begin{bmatrix} \mathbf{c}_{sk}^{b} \\ \vdots \\ \mathbf{c}_{pk}^{b} \end{bmatrix} \right]. \tag{17}$$

Note that the scores $\mathbf{c}_{si}^{b}$ are equal for each between-condition component s at each condition-level i. For convenience, the raw data reproduced from the first component are considered. This yields

$$\mathbf{x}_1^{*} = \left[\mathbf{a}_{1i}^{v}\mathbf{c}_{1i}^{v}, \quad \cdots \quad , \mathbf{a}_{1k}^{v}\mathbf{c}_{1k}^{v} \right] + \mathbf{a}_1^{b} \left[\mathbf{c}_{1i}^{b}, \quad \cdots \quad , \mathbf{c}_{1k}^{b} \right]. \tag{18}$$

It follows from $\mathbf{a}_{1i}^{v} = \mathbf{a}_1^{b}, \cdots, \mathbf{a}_{1k}^{v} = \mathbf{a}_1^{b}$ that $\mathbf{c}_1 = [\mathbf{c}_{1i}^{v} + \mathbf{c}_{1i}^{b}, \quad \cdots \quad , \mathbf{c}_{1k}^{v} + \mathbf{c}_{1k}^{b}]$ and that $\mathbf{x}_1^{*} = \mathbf{a}_1^{b}\mathbf{c}_1$. Thus, it is possible that only a subset of the within-condition loading matrices and between-condition loading matrices is identical and that this subset of components combines within- and between-condition variance. When there is only one between-condition component, i.e., $q^{v} = p > q^{b} = 1$, Equation 17 can be written as

$$\mathbf{x} = \left[\begin{bmatrix} & & \\ [\mathbf{a}_{si}^{v}, \cdots, \mathbf{a}_{pi}^{v}] & & \end{bmatrix} \begin{bmatrix} \mathbf{c}_{si}^{v} \\ \vdots \\ \mathbf{c}_{pi}^{v} \end{bmatrix}, \quad \cdots \quad , [\mathbf{a}_{sk}^{v}, \cdots, \mathbf{a}_{pk}^{v}] \begin{bmatrix} \mathbf{c}_{sk}^{v} \\ \vdots \\ \mathbf{c}_{pk}^{v} \end{bmatrix} \right] + \mathbf{a}_1^{b} \left[\mathbf{c}_{1i}^{b}, \quad \cdots \quad , \mathbf{c}_{1k}^{b} \right]. \tag{19}$$

Theorem 3 describes constraints for the loadings that are compatible with a model combining a single between-condition component with the first within-condition component.

Theorem 3. *If* $q^{v} = p > q^{b} = 1$, *and* $\mathbf{a}_{1i}^{v} = \mathbf{a}_1^{b}, \cdots, \mathbf{a}_{1k}^{v} = \mathbf{a}_1^{b}$ *and* $\mathbf{a}_{si}^{v} \neq \mathbf{a}_1^{b}, \cdots, \mathbf{a}_{sk}^{v} \neq \mathbf{a}_1^{b}$,

then $\mathbf{c}_1 = \left[\mathbf{c}_{1i}^{v} + \mathbf{c}_{1i}^{b}, \quad \cdots \quad , \mathbf{c}_{1k}^{v} + \mathbf{c}_{1k}^{b} \right]$ *and* $\mathbf{c}_s \neq \left[\mathbf{c}_{si}^{v} + \mathbf{c}_{si}^{b}, \quad \cdots \quad , \mathbf{c}_{sk}^{v} + \mathbf{c}_{sk}^{b} \right]$, *for* $s = 2, ..., p$.

Proof. For $\mathbf{a}_{1i}^{v} = \mathbf{a}_1^{b}, \cdots, \mathbf{a}_{1k}^{v} = \mathbf{a}_1^{b}$ Equation 18 can be written as

$$\begin{aligned} \mathbf{x} &= \left[\begin{bmatrix} & & \\ [\mathbf{a}_1^{b}, \mathbf{a}_{si}^{v}, \cdots, \mathbf{a}_{pi}^{v}] & & \end{bmatrix} \begin{bmatrix} \mathbf{c}_{1i}^{v} \\ \mathbf{c}_{si}^{v} \\ \vdots \\ \mathbf{c}_{pi}^{v} \end{bmatrix}, \quad \cdots \quad , [\mathbf{a}_1^{b}, \mathbf{a}_{sk}^{v}, \cdots, \mathbf{a}_{pk}^{v}] \begin{bmatrix} \mathbf{c}_{1k}^{v} \\ \mathbf{c}_{sk}^{v} \\ \vdots \\ \mathbf{c}_{pk}^{v} \end{bmatrix} \right] + \mathbf{a}_1^{b} \left[\mathbf{c}_{1i}^{b}, \quad \cdots \quad , \mathbf{c}_{1k}^{b} \right] \\ &= \left[\begin{bmatrix} & & \\ [\mathbf{a}_{si}^{v}, \cdots, \mathbf{a}_{pi}^{v}] & & \end{bmatrix} \begin{bmatrix} \mathbf{c}_{si}^{v} \\ \vdots \\ \mathbf{c}_{pi}^{v} \end{bmatrix}, \quad \cdots \quad , [\mathbf{a}_{sk}^{v}, \cdots, \mathbf{a}_{pk}^{v}] \begin{bmatrix} \mathbf{c}_{sk}^{v} \\ \vdots \\ \mathbf{c}_{pk}^{v} \end{bmatrix} \right] + \mathbf{a}_1^{b}\mathbf{c}_1, \\ \text{with } \mathbf{c}_1 &= \left[\mathbf{c}_{1i}^{v} + \mathbf{c}_{1i}^{b}, \quad \cdots \quad , \mathbf{c}_{1k}^{v} + \mathbf{c}_{1k}^{b} \right], \text{ for } s = 2, ..., p. \end{aligned} \tag{20}$$

This completes the proof. □

The identity of the loading patterns of the first unrotated within- and between-condition components is a sufficient constraint for the allocation of the between- and within-condition variance on a common component. Theorem 4 describes a somewhat relaxed constraint that is based on an identical shape of the loadings of the first within- and between-condition components but allows for a different scale.

Theorem 4. *If* $q^{v} = p > q^{b} = 1$, *and* $\theta_i\mathbf{a}_{1i}^{v} = \mathbf{a}_1^{b}, \cdots, \theta_k\mathbf{a}_{1k}^{v} = \mathbf{a}_1^{b}$ *and* $\theta_i\mathbf{a}_{si}^{v} \neq \mathbf{a}_1^{b}, \cdots, \theta_k\mathbf{a}_{sk}^{v} \neq \mathbf{a}_1^{b}$,

then $\mathbf{c}_s \neq \left[\theta_i\mathbf{c}_{si}^{v} + \mathbf{c}_{si}^{b}, \quad \cdots \quad , \theta_k\mathbf{c}_{sk}^{v} + \mathbf{c}_{sk}^{b} \right]$, *for* $s = 2, ..., p$ *and* $\theta_i > 0, \cdots, \theta_k > 0$. *and* $\mathbf{c}_1 = \left[\theta_i\mathbf{c}_{1i}^{v} + \mathbf{c}_{1i}^{b}, \quad \cdots \quad , \theta_k\mathbf{c}_{1k}^{v} + \mathbf{c}_{1k}^{b} \right]$

Proof. For $\theta_i\mathbf{a}_{1i}^{v} = \mathbf{a}_1^{b}, \cdots, \theta_k\mathbf{a}_{1k}^{v} = \mathbf{a}_1^{b}$ Equation 18 can be written as

$$\mathbf{x} = \left[\begin{array}{c} [\theta_i \mathbf{a}_1^b, \mathbf{a}_{si}^v, \cdots, \mathbf{a}_{pi}^v] \begin{bmatrix} \mathbf{c}_{1i}^v \\ \mathbf{c}_{si}^v \\ \vdots \\ \mathbf{c}_{pi}^v \end{bmatrix}, \quad \cdots \quad , [\theta_k \mathbf{a}_1^b, \mathbf{a}_{sk}^v, \cdots, \mathbf{a}_{pk}^v] \begin{bmatrix} \mathbf{c}_{1k}^v \\ \mathbf{c}_{sk}^v \\ \vdots \\ \mathbf{c}_{pk}^v \end{bmatrix} \end{array} \right] + \mathbf{a}_1^b \left[\mathbf{c}_{1i}^b, \quad \cdots \quad , \mathbf{c}_{1k}^b \right] \tag{21}$$

$$= \left[[\mathbf{a}_{si}^v, \cdots, \mathbf{a}_{pi}^v] \begin{bmatrix} \mathbf{c}_{si}^v \\ \vdots \\ \mathbf{c}_{pi}^v \end{bmatrix}, \quad \cdots \quad , [\mathbf{a}_{sk}^v, \cdots, \mathbf{a}_{pk}^v] \begin{bmatrix} \mathbf{c}_{sk}^v \\ \vdots \\ \mathbf{c}_{pk}^v \end{bmatrix} \right] + \mathbf{a}_1^b \mathbf{c}_1 ,$$

$$with \ \mathbf{c}_1 = \left[\theta_i \mathbf{c}_{1i}^v + \mathbf{c}_{1i}^b, \quad \cdots \quad , \theta_k \mathbf{c}_{1k}^v + \mathbf{c}_{1k}^b \right], \ for \ s = 2, ..., p.$$

This completes the proof. □

Theorem 4 shows that condition-specific loading patterns that have the same shape, but a different scale are compatible with a model where a single between-condition component is unambiguously allocated on a single within-condition component.

4. Discussion

According to Wood and McCarthy (1984) misallocation of variance occurs when a single between-condition effect that can in principle be allocated on a single PCA component is allocated on more than one component in a given PCA solution. The present study describes constraints that are to be imposed on the component loading matrices in order to avoid misallocation of variance. The following conclusions can be drawn: When a single between-condition effect is allocated on a single component of an initial PCA solution, any rotation of these initial components will result in a misallocation of variance (Theorem 1). This is an algebraic demonstration of what has been discussed elsewhere (Scharf & Nestler, in press; Dien, 2010; Beauducel & Leue, 2015; Beauducel, 2018), namely that, at the level of combined within- and between-condition components, the misallocation of variance is directly related to component rotation. However, component rotation can only result in an optimal allocation of between-condition variance when such a rotational solution exists for a given data set.

Since it has been proposed to perform separate PCAs at each level of a condition factor (Barry et al., 2016), the consequences of this procedure for misallocation of variance were explored. When a PCA is calculated at each level of a condition factor and when a PCA is calculated for a single between-condition factor, an unambiguous allocation of the between-condition variance on a single component combining within- and between-condition variance is possible when the within-condition component loadings have the same shape, even when their scale is different (Theorem 3 and 4). Thus, only when the constraints of Theorem 3 and 4 hold for a given data set, it would be possible to find the solution with optimal allocation of between-condition variance by means of component rotation.

Theorem 3 and 4 also imply that separate PCAs at each level of a condition-factor are not necessarily a way to avoid or eliminate misallocation of variance. When different loading shapes occur at each level of a condition factor in separate PCAs, this indicates that misallocation of variance would occur when the separate components are combined into within- and between variance components. In contrast, when the loading shape is similar in the different PCAs with larger or smaller loadings at each level of the condition factor, the components can be combined into within- and between-components without misallocation of variance.

Finally, it follows from Theorem 4 that perfect congruence coefficients (Tucker, 1951; Wrigley & Neuhaus, 1955) of the loadings of respective components at different levels of the condition factor are not a necessary condition for optimal variance allocation because congruence coefficients also refer to the similarity of the loading magnitude. For optimal variance allocation, a perfect Pearson correlation of the loadings of the respective components at different levels of the condition factor would be necessary.

References

Barry, R. J., De Blasio, F. M., Fogarty, J. S., & Karamacoska, D. (2016). ERP Go/NoGo condition effects are better detected with separate PCAs. *International Journal of Psychophysiology, 106*, 50-64. http://dx.doi.org /10.1016/j.ijpsycho.2016.06.003

Beauducel, A., & Debener, S. (2003). Misallocation of variance in event-related potentials: simulation studies on the effects of test power, topography, andbaseline-to-peak versus principal component quantifications. *Journal of Neuroscience Methods, 124*, 103–112. http://dx.doi.org/10.1016/S0165-0270(02)00381-3.

Beauducel, A. (2018). Recovering Wood and McCarthy's ERP-prototypes by means of ERP-specific Procrustes-rotation. *Journal of Neuroscience Methods, 295*, 20-36. https://doi.org/10.1016 /j.jneumeth.2017.11.011

Beauducel, A., & Leue, A. (2015). Controlling for experimental effects in event-related potentials by means of principal component rotation. *Journal of Neuroscience Methods, 239*, 139–147. http://dx.doi.org /10.1016/j.jneumeth.2014.10.008

Dien, J. (2010). Evaluating two-step PCA of ERP data with Geomin, Infomax, Oblimin, Promax, and Varimax rotations. *Psychophysiology, 47*, 170–183. https://doi.org/10.1111/j.1469-8986.2009.00885.x

Dien J., Khoe, W., & Mangun, G. R. (2007). Evaluation of PCA and ICA of simulated ERPs: Promax vs. Infomax rotations. *Human Brain Mapping, 28*, 742–763. https://doi.org/10.1002/hbm.20304

Harman, H. H. (1976). *Modern factor analysis (3rd ed.).* Chicago, IL: University of Chicago Press.

Jolliffe, I. T., & Cadima J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions Royal Society A, 374*: 20150202. http://dx.doi.org/10.1098/rsta.2015.0202

Kayser, J., & Tenke, C. E. (2003). Optimizing principal component analysis PCA methodology for ERP component identification and measurement: theoretical rationale and empirical evaluation. *Psychophysiology, 114*, 2307–2325. https://doi.org/10.1016/S1388-2457(03)00241-4

Kayser, J., & Tenke, C. E. (2005). Trusting in or breaking with convention: towards a renaissance of principal component analysis in electrophysiology. *Clinical Neurophysiology, 116*, 1747–1753. https://doi.org/ 10.1016/j.clinph.2005.03.020

Scharf, F., & Nestler, S. (in press). Principles behind variance misallocation in temporal exploratory factor analysis for ERP data: Insights from an inter-factor covariance decomposition. *International Journal of Psychophysiology.* https://doi.org/10.1016/j.ijpsycho.2018.03.019

Tucker, L. R. (1951). A Method for Synthesis of Factor Analysis Studies *(Personnel Research Section Report No. 984).* Department of the Army, Washington, DC.

Wood, C. C., & McCarthy, G. (1984). Principal component analysis of event-related potentials: simulation studies demonstrate misallocation of variance across components. *Electroencephalography and Clinical Neurophysiology, 59*, 249–260. https://doi.org/10.1016/0168-5597(84)90064-9

Wrigley, C. S., & Neuhaus, J. O. (1955). The matching of two sets of factors. *American Psychologist, 10*, 418–419.

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The Two-Parameter Odd Lindley Weibull Lifetime Model with Properties and Applications

Jehhan. A. Almamy[1], Mohamed Ibrahim[2], M. S. Eliwa[3], Saeed Al-mualim[1,4] & Haitham M. Yousof[5]

[1] Management Information System Department, Taibah University, Saudi Arabia

[2] Department of Applied Statistics and Insurance, Faculty of Commerce, Damietta University, Damietta, Egypt

[3] Department of Mathematics, Faculty of Science, Mansoura University, Egypt

[4] Department of Statistics, Sana'a University, Yemen

[5] Department of Statistics, Mathematics and Insurance, Benha University, Benha, Egypt

Correspondence: Haitham M. Yousof, Department of Statistics, Mathematics and Insurance, Benha University, Egypt.

Received: March 19, 2018 Accepted: March 27, 2018 Online Published: June 19, 2018

doi:10.5539/ijsp.v7n4p57 URL: https://doi.org/10.5539/ijsp.v7n4p57

Abstract

In this work, we study the two-parameter Odd Lindley Weibull lifetime model. This distribution is motivated by the wide use of the Weibull model in many applied areas and also for the fact that this new generalization provides more flexibility to analyze real data. The Odd Lindley Weibull density function can be written as a linear combination of the exponentiated Weibull densities. We derive explicit expressions for the ordinary and incomplete moments, moments of the (reversed) residual life, generating functions and order statistics. We discuss the maximum likelihood estimation of the model parameters. We assess the performance of the maximum likelihood estimators in terms of biases, variances, mean squared of errors by means of a simulation study. The usefulness of the new model is illustrated by means of two real data sets. The new model provides consistently better fits than other competitive models for these data sets. The Odd Lindley Weibull lifetime model is much better than Weibull, exponential Weibull, Kumaraswamy Weibull, beta Weibull, and the three parameters odd lindly Weibull with three parameters models so the Odd Lindley Weibull model is a good alternative to these models in modeling glass fibres data as well as the Odd Lindley Weibull model is much better than the Weibull, Lindley Weibull transmuted complementary Weibull geometric and beta Weibull models so it is a good alternative to these models in modeling time-to-failure data.

Keywords: Lindley distribution, Weibull distribution, Maximum likelihood, Moments, Order statistics

1. Introduction

The goal of this paper is to introduce a new two parameter alternative to the Weibull, beta Weibull, Lindley Weibull, exponential Weibull, Kumaraswamy Weibull, transmuted complementary Weibull geometric and the tree parameters Odd lindly Weibull (OLW) models that overcomes these mentioned drawbacks.

The probability density function (PDF) and CDF of the Weibull (W) distribution are given by (for $x \geq 0$)

$$g(x,\beta) = \beta x^{\beta-1} \exp\left(-x^{\beta}\right), \tag{1}$$

and

$$G(x,\beta) = 1 - \exp\left(-x^{\beta}\right), \tag{2}$$

respectively, where $\beta > 0$ is a shape parameter. Some useful generalization of the Weibull distribution studied in the literature includes, but are not limited to, Mudholkar and Srivastava (1993), Mudholkar et al. (1995), Mudholkar et al. (1996), Xie and Lai (1995), Ghitany et al. (2005), Famoye et al. (2005), Sarhan and Zaindin (2009), Silva et al. (2010), Aryal and Tsokos (2011), Xie et al. (2002), Lai et al. (2003), Cordeiro et al. (2010), Provost et al. (2011), Cordeiro et al. (2012), Shahbaz et al. (2012), Khan and King (2013), Cordeiro et al. (2013), Merovci and Elbatal (2013), Hanook et al. (2013), Yousof et al. (2015), Cordeiro et al. (2014), Lee et al. (2007), Elbatal and Aryal (2013), Aryal and Elbatl (2015), Afify et al. (2016), Nofal et al. (2016), El-Bassiouny et al. (2016), Yousof et al. (2017a,b,c,d), Aryal et al. (2017a,b), Korkmaz et al. (2017), El-Bassiouny et al. (2017), Alizadeh et al. (2017a,b), Brito et al. (2015). Alizadeh et al. (2018), Yousof et al. (2018), Cordeiro et al. (2018), Hamedani et al. (2018), among others. A state-of-the-art survey on the class of such generalized Weibull distributions can be found in Lai et al. (2001) and Nadarajah (2009)

The probability density function (PDF) and CDF of the OL-G family of distribution (Silva et al. (2017)) are given by

$$f(x;\alpha,\xi) = \alpha^2 (1+\alpha)^{-1} g(x;\xi) \overline{G}(x;\xi)^{-3} \exp\left[-\alpha G(x;\xi) / \overline{G}(x;\xi)\right], \tag{3}$$

and

$$F(x;\alpha,\xi) = 1 - \left[\alpha + \overline{G}(x;\xi)\right](1+\alpha)^{-1}\overline{G}(x;\xi)^{-1}\exp\left[-\alpha G(x;\xi)/\overline{G}(x;\xi)\right], \tag{4}$$

respectively. To this end, by using equations (1), (2) and (3) to obtain the two-parameter OLW PDF (for $x \geq 0$). A random variable X is said to have the OLW distribution if its density function and CDF are given by

$$f(x;\alpha,\beta) = \alpha^2(1+\alpha)^{-1}\beta x^{\beta-1}\left[\exp\left(2x^{\beta}\right)\right]\exp\left[-\alpha\frac{1-\exp\left(-x^{\beta}\right)}{\exp\left(-x^{\beta}\right)}\right], \tag{5}$$

and

$$F(x;\alpha,\beta) = 1 - (1+\alpha)^{-1}\left[\alpha + \exp\left(-x^{\beta}\right)\right]\exp\left(x^{\beta}\right)\exp\left\{-\alpha\left[\exp\left(x^{\beta}\right) - 1\right]\right\}, \tag{6}$$

respectively, we write $X \sim$OLW(α,β), where α is a positive shape parameter. The PDF in (5) and the CDF in (6) are firstly introduced by Silva et al. (2017). Henceforth, the PDF of X in (5) can be easily expressed as

$$f(x) = \sum_{i,k=0}^{\infty} \upsilon_{i,k}\, g\left(x;(i+k+1),\beta\right), \tag{7}$$

where

$$\upsilon_{i,k} = (-1)^k \alpha^{2+k}(\alpha+1)^{-1}\left[(i+k+1)\,i!\right]^{-1}\Gamma(i+k+3)/\Gamma(k+3),$$

and $g(x;\delta,\beta)$ is PDF of Exp-W model with positive parameters δ and β. A handbook, by Rinne (2009), covers the Weibull model in many of its aspects. The study of the family of Exp-W models and their applications attracted the interest of researchers in the nineties. Such interest is growing since then. The CDF of X can be given by integrating (7) as

$$F(x) = \sum_{i,k=0}^{\infty} \upsilon_{i,k}\, G\left(x;(i+k+1),\beta\right), \tag{8}$$

where $G(x;\delta,\beta)$ is PDF of Exp-W model with positive parameters δ and β. For further information about the Exp-W distribution we refer to Mudholkar and Srivastava (1993) and Nadarajah and Kotz (2006). For more details about the OL-G family and its properties see Silva et al. (2017).

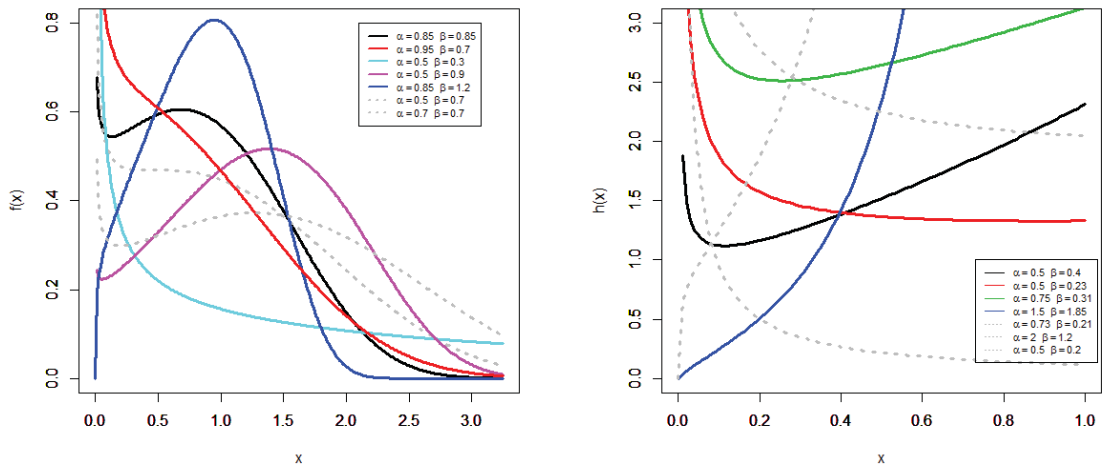


Figure 1. Plots of the OLW PDF and HRF for some parameter values.

The justification for the practicality of the OLW lifetime model is based on the wider use of the W model. Aswell as we are motivated to introduce the OLW lifetime model because it exhibits increasing, decreasing as well as bathtub hazard rates as illustrated in Figure 2. It is shown in Section 1 that the OLW lifetime model can be viewed as a mixture of the

two-parameter Exp-W distributions introduced by Mudholkar and Srivastava (1993) and Mudholkar et al. (1995). It can be viewed as a suitable model for fitting the unimodal and left skewed data. The OLW lifetime model is much better than Weibull, exponential Weibull, Kumaraswamy Weibull, beta Weibull, and the three parameters Odd lindly Weibull with three parameters models so the OLW lifetime model is a good alternative to these models in modeling glass fibres data as well as the OLW lifetime model is much better than the Weibull, Lindley Weibull transmuted complementary Weibull geometric and beta Weibull models so the OLW lifetime model is a good alternative to these models in modeling time-to-failure data.

2. Statistical Properties

2.1 Quantile Functions

Let X be an arbitrary random variable (r.v.) with CDF $F(x;\alpha,\beta)$. For any $u \in (0.1)$, the uth quantile function (QF) $Q(u)$ of the r.v. X is the solution of $u = F(Q(u))$ for all $Q(u) > 0$, from Equation (6), we get

$$(u-1)(1+\alpha)\exp(1+\alpha) = -\frac{1+\alpha-G(Q(u))}{1-G(Q(u))}\exp\left\{-\frac{1+\alpha-G(Q(u))}{1-G(Q(u))}\right\},$$

where

$$-\frac{1+\alpha-G(Q(u))}{1-G(Q(u))},$$

is the Lambert $W(\cdot)$ function of the real argument $(u-1)(1+\alpha)\exp(1+\alpha)$. From Silva et al. (2017), we can write the following equation for QF of the OLW model

$$Q(u) = \left[1-\log\left(1-\left\{1+\alpha\left[1+W_{-1}\left((u-1)(1+\alpha)\exp(1+\alpha)\right)^{-1}\right]\right\}\right)\right]^{\frac{1}{\beta}},$$

where $W(\cdot)$ is Lambert function.

2.2 Moments

The r^{th} ordinary moment of X is given by $\mu'_r = \int_0^\infty x^r f(x)\,dx = E(X^r)$. Using (7), we get

$$\mu'_r = \sum_{i,j,k=0}^{\infty} \upsilon_{i,k}\Upsilon_j^{(i+k+1)}\prod_{m=0}^{\frac{r}{\beta}-1}\left(\frac{r}{\beta}-m\right),\ \forall\ r > -\beta, \tag{9}$$

where

$$\Upsilon_j^{(\omega,r)} = \omega(-1)^i(j+1)^{-(r+\beta)/\beta}\binom{\omega-1}{j},$$

$$\prod_{m=0}^{v-1}(v-m) = \Gamma(1+v) = v(v-1)(v-2)...1, v \in \mathbb{R}^+$$

and $\Gamma(\zeta) = \int_0^\infty x^{\zeta-1}e^{-t}dx$ is the complete gamma function. The r^{th} incomplete moment of X, say $\varphi_r(t)$, is given by $\varphi_r(t) = \int_0^t x^r f(x)\,dx$. Using Equation (7), we obtain

$$\varphi_r(t) = \gamma\left(1+r\beta^{-1}, t^{-\beta}\right)\sum_{i,j,k=0}^{\infty}\upsilon_{i,k}\Upsilon_j^{(i+k+1,r)},\ \forall\ r > -\beta, \tag{10}$$

where $\gamma(\zeta,x) = \int_0^x x^{\zeta-1}e^{-x}dx$ is the incomplete gamma function.

2.3 Order Statistics and Their Moments

Let $X_1,\ldots,X_n$ be a random sample from the OLW model of distributions and let $X_{1:1},\ldots,X_{n:n}$ be the corresponding order statistics. The PDF of the i^{th} order statistic, say $X_{i:n}$, can be expressed as

$$f_{i:n}(x) = [\mathrm{B}(i,n-i+1)]^{-1} f(x) F(x)^{i-1}\ [1-F(x)]^{n-i}, \tag{11}$$

where $B(\cdot,\cdot)$ is the beta function. Substituting (5) and (6) in Equation (11), we obtain

$$f_{i:n}(x) = \sum_{m,p=0}^{\infty}\sum_{j=0}^{k+n-i}\upsilon_{j,m,p}\, g(x;(j+m+p+1),\beta),$$

where

$$\upsilon_{j,m,h} = \sum_{k=0}^{i-1} (-1)^{k+m} \alpha^{j+m+2} (1+\alpha)^{-(j+1)} [m!(j+m+p+1)]^{-1} \times [\mathrm{B}(i, n-i+1)]^{-1} \binom{j+m+p}{j+m} \binom{k+n-1}{j} \binom{i-1}{k}.$$

Then, the q^{th} moment of $X_{i:n}$ is given by

$$E\left(X_{i:n}^{q}\right) = \sum_{m,p,h=0}^{\infty} \sum_{j=0}^{k+n-i} \upsilon_{j,m,p} \, \Upsilon_{h}^{(j+m+p+1,q)} \prod_{w=0}^{\frac{q}{\beta}-1} \left(\frac{q}{\beta} - w\right), \ \forall \ q > -\beta. \tag{12}$$

Based upon the moments in Equation (12), we can derive explicit expressions for the L-moments of X as infinite weighted linear combinations of the means of suitable OLW order statistics. They are linear functions of expected order statistics defined by

$$\lambda_r = r^{-1} \sum_{d=0}^{r-1} (-1)^d \binom{r-1}{d} E\left(X_{r-d:r}\right), \ r \geq 1.$$

2.4 Moment of Residual and Reversed Residual Lifes

The n^{th} moment of the residual life, say $z_n(t) = E[(X-t)^n \mid X > t]$, $n = 1, 2, \ldots$, uniquely determines $F(x)$. The n^{th} moment of the residual life of X is given by

$$z_n(t) = \frac{\int_t^{\infty} (x-t)^n dF(x)}{1-F(t)}.$$

We can write

$$\begin{aligned} z_n(t) &= [1-F(t)]^{-1} \sum_{i,k=0}^{\infty} \sum_{r=0}^{n} \upsilon_{i,k} (-t)^{n-r} \binom{n}{r} \int_t^{\infty} x^r \, g\left(x; (i+k+1), \beta\right) dx \\ &= \gamma\left(1 + n\beta^{-1}, t^{-\beta}\right) [1-F(t)]^{-1} \sum_{i,j,k=0}^{\infty} \sum_{r=0}^{n} \Upsilon_{i,j,k,r}^{(i+k+1,n)}, \ \forall \ n > -\beta, \end{aligned}$$

where

$$\Upsilon_{i,j,k,r}^{(i+k+1,n)} = \upsilon_{i,k} t^{n-r} (i+k+1) (-1)^{i+n-r} (j+1)^{-(n+\beta)/\beta} \binom{i+k}{j} \binom{n}{r}.$$

The n^{th} moment of the reversed residual life, say $Z_n(t) = E\left[(t-X)^n \mid X \leq t\right]$, for $t > 0$ and $n = 1, 2, \ldots$, uniquely determines $F(x)$. We have

$$Z_n(t) = \frac{\int_0^t (t-x)^n dF(x)}{F(t)}.$$

Then, the n^{th} moment of the reversed residual life of X becomes

$$\begin{aligned} Z_n(t) &= F(t)^{-1} \sum_{i,k=0}^{\infty} \sum_{r=0}^{n} \upsilon_{i,k} (-1)^r \binom{n}{r} t^{n-r} \int_0^t x^r \, g\left(x; (i+k+1), \beta\right) dx \\ &= \gamma\left(1 + n\beta^{-1}, t^{-\beta}\right) F(t)^{-1} \sum_{i,j,k=0}^{\infty} \sum_{r=0}^{n} \Omega_{i,j,k,r}^{(i+k+1,n)}, \ \forall \ n > -\beta, \end{aligned}$$

where

$$\Omega_{i,j,k,r}^{(i+k+1,n)} = \upsilon_{i,k} t^{n-r} (i+k+1) (-1)^{i+r} (j+1)^{-(n+\beta)/\beta} \binom{i+k}{j} \binom{n}{r}.$$

3. Maximum Likelihood Method

We consider the estimation of the unknown parameters of the OLW model from complete samples only by maximum likelihood method. The MLEs of the parameters of the OLW (α,β) model is now discussed. Let $x_1,\ldots,x_n$ be a random sample of this distribution with parameter vector $\mathbf{\Psi} = (\alpha,\beta)^{\intercal}$. The log-likelihood function for $\mathbf{\Psi}$, say $\ell = \ell(\mathbf{\Psi})$, is given by

$$\ell = \ell(\mathbf{\Psi}) = 2n\log(\alpha) - n\log(1+\alpha) + n\log\beta + (\beta-1)\sum_{i=1}^{n}\log(x_i) + 2\sum_{i=1}^{n}x_i^{\beta} - \alpha\sum_{i=1}^{n}\frac{1-\exp\left(-x_i^{\beta}\right)}{\exp\left(-x_i^{\beta}\right)},$$

the last equation can be maximized either by using the different programs like R (`optim` function), SAS (`PROC NLMIXED`) or by solving the nonlinear likelihood equations obtained by differentiating Equation 13. The score vector elements, $\mathbf{U}(\mathbf{\Psi}) = \frac{\partial\ell}{\partial\mathbf{\Psi}} = \left(\frac{\partial\ell}{\partial\alpha},\frac{\partial\ell}{\partial\beta}\right)^{\intercal}$ can be easely obtained, we can obtain the estimates of the unknown parameters by setting the score vector to zero, $\mathbf{U}(\widehat{\mathbf{\Psi}}) = \mathbf{0}$. Solving these equations simultaneously gives the MLEs $\widehat{\alpha}$ and $\widehat{\beta}$. For the OLW distribution, all the second order derivatives exist. The interval estimation of the model parameters requires the 2×2 observed information matrix $J(\mathbf{\Psi}) = \{J_{ij}\}$ for $i,j = \alpha,\beta$. The multivariate normal $N_2(0, J(\widehat{\mathbf{\Psi}})^{-1})$ distribution, under standard regularity conditions, can be used to provide approximate confidence intervals for the unknown parameters, where $J(\widehat{\mathbf{\Psi}})$ is the total observed information matrix evaluated at $\widehat{\mathbf{\Psi}}$. Then, approximate $100(1-\delta)\%$ confidence intervals for α and β can be determined by: $\widehat{\alpha} \pm z_{\delta/2}\sqrt{\widehat{J}_{\alpha\alpha}}$ and $\widehat{\beta} \pm z_{\delta/2}\sqrt{\widehat{J}_{\beta\beta}}$, where $z_{\delta/2}$ is the upper δ^{th} percentile of the standard normal model. Further works could be addressed using different methods to estimate the OLW parameters such as least squares, moments, weighted least squares, Jackknife, Cram'er-von-Mises, bootstrap, Bayesian analysis, Anderson-Darling, among others, and compare the estimators based on these methods.

4. Simulation Studies

We consider a random sample of size n = 50, 150, 200, 250, 300 and 500 from our density corresponding to particular choices of the parameters as follows: α=0.25, β=0.5, α=1.5, β=0.9 and α=0.8, β=1.5, the results are presented in Tables 1, 2 and 3 respectively. Below we provide the MLEs, biases (Bias), variances (Var), mean square of errors (MSEs) and Confidence Interval for the estimates of all the parameters under both the methods of estimation. The log-likelihood function can be maximized directly via the R-package or by solving the nonlinear likelihood equations obtained by differentiating the PDF (5) (using the optim function as well as the Max-BFGS subroutines. One can observe the estimates of the unknown parameters by setting the score vector to zero, and then using any statistical software to solve them numerically. The results show that the maximum likelihood estimation performs well. In general, the biases, variances and MSEs of the parameters are reasonably small. The biases, variances and MSEs always decrease as the sample size increases. The results suggest that the maximum likelihood method can be used to estimate the parameters of the OLW model.

Table 1 provides the biases, variances, MSEs and Confidence Interval under the method of maximum likelihood. We consider 1000 simulations for drawing random samples each of size $n = 50, 150, 200, 250, 300$ and 500 drawn from our density respectively when $\alpha = 0.25$ and $\beta = 0.5$.

Table 1. The MLEs, Bias, Var, and MSE values for the OLWD

n	Parameter (MLE)	Bias	Var	MSE	Confidence Interval
50	α=0.25(0.247229)	-0.002771	0.0017944	0.001802	(0.1691,0.3366)
	β=0.5(0.507311)	0.007311	0.0011746	0.001228	(0.4418,0.5759)
150	α=0.25(0.247489)	-0.002511	0.0006301	0.000636	(0.1993,0.2985)
	β=0.5(0.503367)	0.003367	0.0003829	0.000394	(0.4677,0.5436)
200	α=0.25(0.248688)	-0.001312	0.0004553	0.000457	(0.2072,0.2918)
	β=0.5(0.502525)	0.002525	0.0002828	0.000289	(0.4699,0.5368)
250	α=0.25(0.249041)	-0.000959	0.0003657	0.000367	(0.2148,0.2886)
	β=0.5(0.501742)	0.001742	0.000216	0.000219	(0.4726,0.5316)
300	α=0.25(0.249389)	-0.000611	0.00032	0.00032	(0.2155,0.2861)
	β=0.5(0.501331)	0.001331	0.0001873	0.000189	(0.4743,0.5301)
500	α=0.25(0.249847)	-0.000153	0.0001874	0.000187	(0.2245,0.2787)
	β=0.5(0.500251)	0.000251	0.0001117	0.000112	(0.4798,0.5218)

Table 2 provides the MLEs, biases, variances, MSEs and Confidence Interval under the method of maximum likelihood. We consider 1000 simulations for drawing random samples each of size $n = 50, 150, 200, 250, 300$ and 500 drawn from our density respectively when α=1.5 and β=0.9.

Table 2. The MLEs, Bias, Var, and MSE values for the OLWD

n	Parameter(MLE)	Bias	Var	MSE	Confidence Interval
50	α = 1.5(1.529613)	0.029613	0.033058	0.033935	(1.2203,1.9334)
	β = 0.9(0.929152)	0.029152	0.0156265	0.016476	(0.7296,1.2245)
150	α = 1.5(1.51163)	0.01163	0.0096352	0.00977	(1.3240,1.7228)
	β = 0.9(0.907027)	0.007027	0.0038567	0.003906	(0.7955,1.0367)
200	α = 1.5(1.508664)	0.008664	0.0077297	0.007805	(1.3531,1.6969)
	β = 0.9(0.906696)	0.006696	0.0031	0.003145	(0.8050,1.0131)
250	α = 1.5(1.503555)	0.003555	0.0056863	0.005699	(1.3631,1.6487)
	β = 0.9(0.905057)	0.005057	0.0023803	0.002406	(0.8178,1.0088)
300	α = 1.5(1.506448)	0.006448	0.0047551	0.004797	(1.3760,1.6528)
	β = 0.9(0.904546)	0.004546	0.0020092	0.00203	(0.8215,0.9957)
500	α = 1.5(1.502812)	0.002812	0.0029378	0.002946	(1.4031,1.6138)
	β = 0.9(0.903687)	0.003687	0.001307	0.001321	(0.8385,0.9740)

Table 3 provides the biases, variances, MSEs and Confidence Interval under the method of maximum likelihood. We consider 1000 simulations for drawing random samples each of size $n = 50, 150, 200, 250, 300$ and 500 drawn from our density respectively when α=0.8 and β=1.5.

Table 3. The MLEs, Bias, Var, and MSE values for the OLWD

n	Parameter (MLE)	Bias	Var	MSE	Confidence Interval
50	α = 0.8(0.80632)	0.00632	0.0091889	0.009229	(0.6242,1.0034)
	β = 1.5(1.538177)	0.038177	0.0255296	0.026987	(1.2617,1.8741)
150	α = 0.8(0.799961)	-0.000039	0.0027371	0.002737	(0.7012,0.9048)
	β = 1.5(1.512985)	0.012985	0.0073598	0.007528	(1.3508,1.6777)
200	α = 0.8(0.800502)	0.000502	0.0020934	0.002094	(0.7089,0.8903)
	β = 1.5(1.510718)	0.010718	0.0058925	0.006007	(1.3640,16650)
250	α = 0.8(0.800997)	0.000997	0.0018525	0.001854	(0.7174,0.8887)
	β = 1.5(1.506141)	0.006141	0.0044795	0.004517	(1.3843,1.6418)
300	α = 0.8(0.800886)	0.000886	0.0014993	0.00150	(0.7275,0.8757)
	β = 1.5(1.505427)	0.005427	0.0037086	0.003738	(1.3875,1.6335)
500	α = 0.8(0.80081)	0.00081	0.000879	0.00088	(0.7418,0.8593)
	β = 1.5(1.503156)	0.003156	0.002423	0.002433	(1.4061,1.5991)

From Tables 1, 2 and 3, we note that the Bias is reduced as the sample size is increased.

5. Real Data Analysis

In this section, we illustrate the empirical importance of the OLW model and other lifetime distributions using two applications to real data.

The first data set (I): The first set consists of 63 observations of the strengths of 1.5 cm glass fibres, originally obtained by workers at the UK National Physical Laboratory. The data are:

0.55, 0.74, 0.77, 0.81, 0.84, 0.93, 1.04, 1.11, 1.13, 1.24, 1.25, 1.27, 1.28, 1.29, 1.30, 1.36, 1.39, 1.42, 1.48, 1.48, 1.49, 1.49, 1.50, 1.50, 1.51, 1.52, 1.53, 1.54, 1.55, 1.55, 1.58, 1.59, 1.60, 1.61, 1.61, 1.61, 1.61, 1.62, 1.62, 1.63, 1.64, 1.66, 1.66, 1.66, 1.67, 1.68, 1.68, 1.69, 1.70, 1.70, 1.73, 1.76, 1.76, 1.77, 1.78, 1.81, 1.82, 1.84, 1.84, 1.89, 2.00, 2.01, 2.24. These data have also been analyzed by Smith and Naylor (1987). For this data set, we shall compare the fits of the OLW distribution with some competitive models like Weibull (W), exponential Weibull (EW), Kumaraswamy Weibull (KwW) (Cordeiro et al., 2010), beta Weibull (BW) (Lee et al., 2007), and Odd lindly Weibull with three parameters (OLW*) (Silva et al. 2016).

The second data set (II): represents 40 observations of time-to-failure ($10^3/h$) of turbocharger of one type of engine, see Xu et al. (2003). The data are: 1.6, 3.5, 4.8, 5.4, 6.0,6.5, 7.0, 7.3, 7.7, 8.0, 8.4, 2.0, 3.9, 5.0, 5.6, 6.1, 6.5, 7.1, 7.3, 7.8, 8.1, 8.4, 2.6, 4.5, 5.1, 5.8, 6.3, 6.7, 7.3, 7.7, 7.9, 8.3, 8.5, 3.0, 4.6, 5.3, 6.0, 8.7, 8.8, 9.0. This data set is used to compare the fits of the OLW lifetime model with some competitive models like W, Lindley Weibull (LiW) (Cordeiro et al. 2018), transmuted complementary Weibull geometric (TCWG) (Afify et al. 2014) and BW models. All parameters of these distribution are positive numbers. In Tables 1 and 2, the MLEs and their standard errors (SEs) (in parentheses) of the parameters from the fitted models and the values of the Akaike Information Criterion (AIC), Cram er-von Mises (W^*) and Anderson-Darling (A^*) goodness-of-fit statistics are presented. According to the lowest values of the AIC, W^* and A^* statistics.

For the first data set, the OLW model provides the best fit. The empirical PDF and CDF for the OLW are displayed in Figures 2 and 3 respectively, which support the results of Table 4.

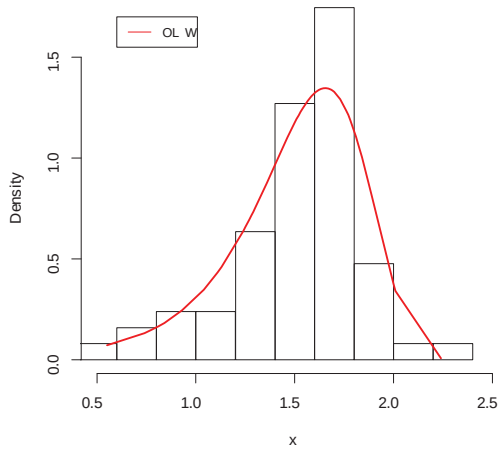


Figure 3. Estimated PDF for data set I.

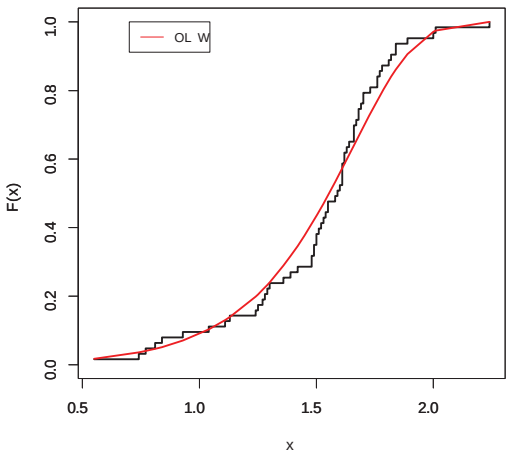


Figure 4. Estimated CDF for data set I.

For the second data set, the OLW model provides the best fit. The empirical PDF and CDF for the OLW are displayed in Figures 4 and 5 respectively, which support the results of Table 2.

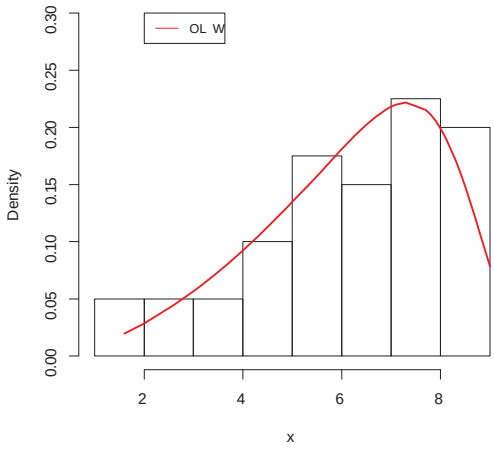


Figure 4. Estimated PDF for data set II.

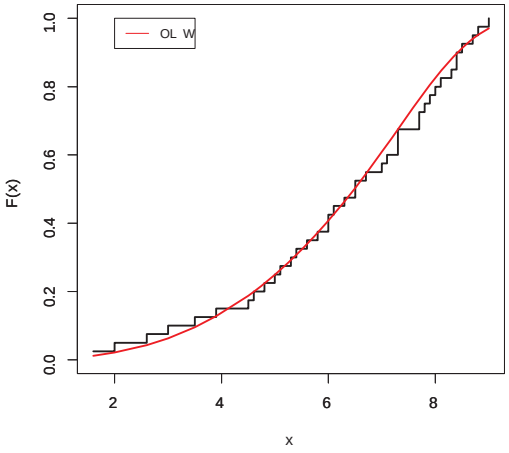


Figure 5. Estimated PDF for data set II.

Hence, we prove empirically that the proposed model provides better fits in two applications than other seven extended Weibull distributions with two, three and four parameters. There are too many models to fit and this fact really shows that the OLW model can be a good alternative for modeling survival data.

6. Discussion

In this work, we propose and study a new two-parameter lifetime model, called the Odd Lindley Weibull (OLW) distribution, which extends the Weibull distribution. The OLW distribution is motivated by the wide use of the Weibull model in many applied areas and also for the fact that this new generalization provides more flexibility to analyze real data. The OLW density function can be written as a linear combination of the exponentiated W (Exp-W) densities. We derive explicit expressions for the ordinary and incomplete moments, moments of the (reversed) residual life, generating functions and order statistics. We discuss the maximum likelihood estimation of the model parameters. We assess the performance of the maximum likelihood estimators in terms of biases, variances, mean squared of errors by means of a simulation study. The usefulness of the new model is illustrated by means of two real data sets. The new model provides consistently better fits than other competitive models for these data sets. The OLW lifetime model is much better than Weibull, exponential Weibull, Kumaraswamy Weibull, beta Weibull, and the three parameters Odd lindly Weibull with three parameters models so the OLW lifetime model is a good alternative to these models in modeling glass fibres data as well as the OLW lifetime model is much better than the Weibull, Lindley Weibull transmuted complementary Weibull geometric and beta Weibull models so the OLW lifetime model is a good alternative to these models in modeling time-to-failure data. We hope that the new distribution will attract wider applications in reliability, engineering and other areas of research. Finally, as a future work we will consider bivariate and multivariate extension of the OLW distribution. In particular with the copula based construction method, trivariate reduction etc.

Table 4. The MLEs(SEs inparentheses) for some fitted models to data set **I** and the AIC,W^* and A^* values

Model	$\hat{\alpha}$	$\hat{b}$	$\hat{\beta}$	$\hat{\lambda}$	AIC	W^*	A^*
W	–	–	5.781	1.628	34.414	0.237	1.304
	–	–	(0.576)	(0.037)			
EW	0.671	–	7.285	1.718	35.351	0.636	3.484
	(0.249)	–	(1.707)	(0.086)			
TW	–	-0.5010	5.1498	0.6458	34.6720	1.0358	0.1691
	–	(0.2741)	(0.6657)	(0.0235)			
OLLW	–	0.9439	6.0256	0.6159	36.3736	1.2364	0.2194
	–	(0.2689)	(1.3478)	(0.0164)			
BW	0.620	10.249	7.759	2.382	37.179	0.196	1.089
	(0.248)	(95.117)	(2.023)	(2.897)			
KwW	0.606	0.214	6.908	1.337	35.252	0.161	0.908
	(0.162)	(0.029)	(0.004)	(0.003)			
OLW*	0.049	–	1.102	0.492	34.387	0.153	0.870
	(0.087)	–	(0.527)	(0.494)			
OLW	0.2026	–	1.716	–	**33.427**	**0.138**	**0.813**
	(0.0317)	–	(0.087)	–			

Table 5. The MLEs(SEs inparentheses) for some fitted models to data set **II** and the AIC,W^* and A^* values

Model	$\hat{\alpha}$	$\hat{b}$	$\hat{\beta}$	$\hat{\lambda}$	AIC	W^*	A^*
W	–	–	3.872	6.920	82.48	0.0769	0.5730
	–	–	(0.517)	(0.294)			
LiW	–	0.898	0.169	3.499	81.89	0.0636	0.4815
	–	(1.093)	(0.073)	(0.633)			
TCWG	0.188	-8.9×10^{-5}	0.2059	2.7881	81.32	0.0496	0.3766
	(0.046)	(0.647)	(0.2747)	(0.8733)			
BW	0.075	11.242	0.240	115.43	79.04	0.0210	0.1696
	(0.030)	(3.850)	(0.102)	(489.0)			
OLW	8.309	–	0.188	–	**77.90**	**0.0159**	**0.1018**
	(0.148)	–	(0.046)	–			

Acknowledgements

The authors are thankful to the Editor and the reviewers for their constructive comments and suggestions which greatly improved the current version.

References

Afify, A. Z., Cordeiro, G. M., Yousof, H. M., Abdus, S., & E. M. M. (2016). *The Marshall-Olkin additive Weibull distribution with variable shapes for the hazard rate.* Hacettepe Journal of Mathematics and Statistics, forthcoming.

Afify, A. Z., Nofal, Z. M., & Butt, N. S. (2014). Transmuted complementary Weibull geometric distribution. *Pak. J. Stat. Oper. Res., 10*, 435-454.

Afify, A. Z., Nofal, Z. M., Yousof, H. M., El Gebaly, Y. M., & Butt, N. S. (2015). The transmuted Weibull Lomax distribution: properties and application. *Pak. J. Stat. Oper. Res., 11*, 135-152.

Alizadeh, M., Ghosh, I., Yousof, H. M., Rasekhi, M., & Hamedani G. G. (2017a). The generalized Odd generalized exponential family of distributions: properties, characterizations and applications. *J. Data Sci. 15*, 443-466.

Alizadeh, M., Lak, F., Rasekhi, M., Ramires, T. G., Yousof, H. M., & Altun, E. (2017b). *The Odd log-logistic Topp Leone G family of distributions: heteroscedastic regression models and applications.* Computational Statistics, forthcoming.

Alizadeh, M., Yousof, H. M. Rasekhi, M., & Altun, E. (2018). *The Odd loglogistic Poisson-G Family of distributions.* Journal of Mathematical Extensions, forthcoming.

Aryal, G. R., & Tsokos, C. P. (2011). Transmuted Weibull distribution: a generalization of the Weibull probability distribution. *European Journal of Pure and Applied Mathematics, 4*, 89-102.

Aryal, G. R., Ortega, E. M., Hamedani, G. G., & Yousof, H. M. (2017a). The Topp-Leone Generated Weibull distribution: regression model, characterizations and applications. *International Journal of Statistics and Probability, 6*, 126-141.

Aryal, G. R., & Yousof, H. M. (2017b). *The exponentiated generalized-G Poisson family of distributions.* Economic Quality Control, forthcoming.

Alizadeh, M., Merovci, F., & Hamedani, G. G. (2015). *Generalized transmuted family of distributions: properties and applications.* Hacettepa Journal of Mathematics and Statistics, forthcomig.

Brito, E., Cordeiro, G. M., Yousof, H. M., Alizadeh, M. & Silva , G. O. (2017). *Topp-Leone Odd Log-Logistic Family of Distributions.* Journal of Statistical Computation and Simulation, 87(15), 3040–3058.

Cordeiro, G. M., Afify, A. Z., Yousof, H. M., Cakmakyapan, S., & Ozel, G. (2018). *The Lindley Weibull distribution: properties and applications.* Anais da Academia Brasileira de Ciências, forthcoming.

Cordeiro, G. M., Yousof, H. M., Ramires, T. G., & Ortega, E. M. M. (2018). The Burr XII system of densities: properties, regression model and applications. *Journal of Statistical Computation and Simulation, 88*(3), 432-456.

Cordeiro, G. M., Ortega, E. M., & da Cunha, D. C. C. (2013). The exponentiated generalized class of distributions. *Journal Data Sci., 11*, 1-27.

Cordeiro, G. M., Ortega, E. M., & Nadarajah, S. (2010). The Kumaraswamy Weibull distribution with application to failure data. *Journal of the Franklin Institute, 347*, 1399-1429.

Cordeiro, G. M., Ortega, E. M., & Silva, G. O. (2012). *The Kumaraswamy modified Weibull distribution: theory and applications. Journal of Statistical Computation and Simulation*, 1-25.

El-Bassiouny, A. H., Medhat, E., Abdelfattah, M., & Eliwa, M. S. (2016). Mixture of exponentiated generalized Weibull-Gompertz distribution and its applications in reliability. *J. Stat. Appl. , 5*(3), 1-14.

El-Bassiouny, A. H., Medhat, E., Abdelfattah, M., & Eliwa, M. S. (2017). Exponentiated generalized Weibull-Gompertz distribution with application in survival analysis. *J. Stat. Appl. Pro., 6*(1), 7-16

Ghitany, M. E., Al-Hussaini, E. K., & Al-Jarallah, R. A. (2005). Marshall-Olkin extended Weibull distribution and its application to censored data. *Journal Applied Statistics, 32*, 1025-1034.

Hamedani G. G. Yousof, H. M., Rasekhi, M., Alizadeh, M., & Najibi, S. M. (2018a). *Type I general exponential class of distributions.* Pak. J. Stat. Oper. Res., forthcoming. Mathematics. forthcoming.

Hamedani G. G. Rasekhi, M., Najibi, S. M., Yousof, H. M., & Alizadeh, M., (2018b). *Type II general exponential class of distributions.* Pak. J. Stat. Oper. Res., forthcoming.

Khan, M. N. (2015). *The modified beta Weibull distribution.* Hacettepe Journal of Mathematics and Statistics, forthcoming.

Khan, M. S., & King, R. (2013). Transmuted modified Weibull distribution: a generalization of the modified Weibull probability distribution. *European Journal of Pure and Applied Mathematics, 6*, 66-88.

Korkmaz, M. C., Yousof, H. M., & Hamedani, G. G. (2017). The exponential Lindley Odd log-logistic G family: properties, characterizations and applications. *Journal of Statistical Theory and Applications,* forthcoming.

Lai, C. D., Xie, M., & Murthy, D. N. P. (2001). *Bathtub-shaped failure rate life distributions.* pages 69–104

Lai, C. D., Xie, M., & Murthy, D. N. P. (2003). *A modified Weibull distribution.* IEEE Transactions on Reliability, 52, 33-37.

Lee E. T., & Wang J. W. (2003). *Statistical Methods for Survival Data Analysis.* 3rd ed.,Wiley, NewYork.

Mead, M. E., & Afify, A. Z. (2016). *On five-parameter Burr XII distribution: properties and applications.* South African Statistical Journal, Forthcomig.

Mudholkar, G. S., & Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure-real data. *IEEE Transactions on Reliability, 42*, 299-302.

Mudholkar, G. S., Srivastava, D. K., & Freimer, M. (1995). The expnentiated Weibull family: a reanalysis of the bus-motor-failure data. *Technometrics, 37*, 436-445.

Mudholkar, G. S., Srivastava, D. K., & Kollia, G. D. (1996). A generalization of the Weibull distribution with application to the analysis of survival data. *Journal of the American Statistical Association, 91*, 1575-1583.

Murthy, D.N.P., Xie, M., & Jiang, R. (2004). *Weibull Models.* 1st ed., John Wiley & Sons, Hoboken, NJ.

Nadarajah, S. (2009). *Bathtub-shaped failure rate functions. Quality and Quantity, 43*, 855–863.

Nadarajah, S., & Kotz, S. (2005). On some recent modifications of Weibull distribution. *IEEE Trans. Reliab., 54*, 561-562.

Nadarajah, S., & Kotz, S. (2006). The exponentiated type distributions. *Acta Applicandae Mathematica, 92*, 97-111.

Nofal, Z. M., Afify, A. Z., Yousof, H. M., & Cordeiro, G. M. (2017). The generalized transmuted-G family of distributions. *Communications in Statistics-Theory and Methods, 46*, 4119-4136.

Nofal, Z. M., Afify, A. Z., Yousof, H. M., Granzotto, D. C. T., & Louzada, F. (2016). Kumaraswamy transmuted exponentiated additive Weibull distribution. *International Journal of Statistics and Probability, 5*, 78-99.

Rinne, H. (2009). *The Weibull distribution.* Chapman & Hall, London.

Shaw, W. T., & Buckley, I. R. C. (2007). The alchemy of probability distributions: beyond Gram-Charlier expansions and a skew-kurtotic-normal distribution from a rank transmutation map. Research report.

Silva, G. O., Ortega, E. M. M., & Cordeiro, G. M. (2010). The beta modified Weibull distribution. *Lifetime Data Analysis, 16*, 409-430.

Silva, F. S., Percontini, A., de Brito, E., Ramos, M. W., Venancio, R., & Cordeiro, G. M. (2017). The Odd Lindley-G Family of Distributions.*Austrian Journal of Statistics, 46*(1), 65-87.

Xie, M., Tang, Y., & Goh, T. N. (2002). A modified Weibull extension with bathtub failure rate function. *Reliability Engineering and System Safety, 76*, 279-285.

Xie, M., & Lai, C. D. (1995). Reliability analysis using an additive Weibull model with bathtub-shaped failure rate function. *Reliability Engineering and System Safety, 52*, 87-93.

Yousof, H. M., Afify, A. Z., Alizadeh, M., Butt, N. S., Hamedani, G. G., & Ali, M. M. (2015). The transmuted exponentiated generalized-G family of distributions. *Pak. J. Stat. Oper. Res., 11*, 441-464.

Yousof, H. M., Afify, A. Z., Alizadeh, M., Nadarajah, S., Aryal, G. R., & Hamedani, G. G. (2017a). *The Marshall-Olkin generalized-G family of distributions with Applications.* Communications in Statistics-Theory and Methods, forthcoming.

Yousof, H. M., Afify, A. Z., Cordeiro, G. M., Alzaatreh, A., & Ahsanullah, M. (2017b). *A new four-parameter Weibull model.* Journal of Statistical Theory and Applications, forthcoming.

Yousof, H. M., Afify, A. Z., Hamedani, G. G., & Aryal, G. (2016). The Burr X generator of distributions for lifetime data. *Journal of Statistical Theory and Applications, 16*, 288C305.

Yousof, H. M., Alizadeh, M., Jahanshahiand, S. M. A., Ramires, T. G., Ghosh, I., & Hamedani G. G. (2017). The transmuted Topp-Leone G family of distributions: theory, characterizations and applications. *Journal of Data Science, 15*, 723-740.

Yousof, H. M., Altun, E., Ramires, T. G., Alizadeh, M., & Rasekhi, M. (2018). A new family of distributions with properties, regression models and applications. *Journal of Statistics and Management Systems, 21*(1), 163-188.

Yousof, H. M. Majumder, M., Jahanshahi, S. M. A., Ali, M. M., & Hamedani G. G. (2017c). *A new Weibull class of distributions: theory, characterizations and applications.* Communications for Statistical Applications and Methods, forthcoming.

Yousof, H. M., Rasekhi, M., Afify, A. Z., Alizadeh, M., Ghosh, I., & Hamedani G. G. (2017d). The beta Weibull-G family of distributions: theory, characterizations and applications. *Pakistan Journal of Statistics, 33*, 95-116.

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The Impact of Sidewalks on Vehicle-Pedestrian Crash Severity

Mehrnaz Doustmohammadi[1], Niloufar Shirani Bidabadi[1], Sumalatha Kesaveraddy[2], Michael Anderson[1]

[1] Department of Civil and Environmental Engineering, The University of Alabama in Huntsville Huntsville, USA

[2] Transportation Engineer at Baslee Engineering Solutions, Tampa, USA

Correspondence: Mehrnaz Doustmohammadi, Department of Civil and Environmental Engineering, The University of Alabama in Huntsville, Huntsville, AL 35899, USA. Tel: 1-205-243-6717. E-mail: md0033@uah.edu

Received: April 18, 2018 Accepted: May 9, 2018 Online Published: June 19, 2018

doi:10.5539/ijsp.v7n4p69 URL: https://doi.org/10.5539/ijsp.v7n4p69

Absrtact

Walking is a sustainable mode of transportation that has several benefits related to improved health and reducing traffic congestion. The drawback to walking as a mode of transportation is the increased potential to be involved in a severe crash, which is greater than when two automobiles are involved in a crash. This paper provides a statistical analysis of pedestrian crashes that occurred in two Alabama cities where the crashes are divided into those where a sidewalk was present and those where a sidewalk was not present. The goal of the paper is to determine the difference in crash experiences and variables that contribute to vehicle-pedestrian crashes associated with the presence of the sidewalk. The paper uses binary logistic regression to develop models of pedestrian crashes and evaluates the models to determine factors that contribute the pedestrian crashes. The paper concludes that pedestrian crashes often happen in the evenings, with low lighting and visibility levels, independent of the presence of sidewalks.

Keywords: pedestrian crash analysis, binary logistic regression, sidewalks

1. Introduction

1.1 Introduction to the Problem

Walking is a vital and sustainable means of transportation and is gaining popularity. The 2009 National Household Travel Survey (NHTS) presented that an estimated 42 billion walking trips are made every year in the US, accounting for 10.5% of the total trips taken (*1*). The safety of these pedestrians therefore is a top priority. In 2014, almost 5,000 pedestrian were killed and 65,000 injured in traffic crashes in the US, with 78% occurring in urban areas (*2*). In Alabama, there were a total of 759 vehicle-pedestrian crashes resulting in 283 fatalities and incapacitating injuries, with another 387 pedestrians injured, with 84% being reported in urban areas (*3*).

To ease pedestrian movements, sidewalks are usual constructed along roadways to allow for those walking a quality, weather restraint surface to make their trip. Additionally, the presence of a sidewalk provides legitimacy to the walking trip and a perceived level safety upon which the pedestrian might use to justify making their trip. However, even with a sidewalk in place, there is still the possibility of a vehicle-pedestrian crash to occur.

This paper provides a statistical analysis of crashes that occurred in two Alabama cities where the pedestrian crashes are divided into those where a sidewalk was present and those where a sidewalk was not present. The goal of the paper is to determine the difference in crash experiences and variables that contribute to vehicle-pedestrian crashes associated with the presence of the sidewalk. The paper uses binary logistic regression to develop models of pedestrian crashes and evaluates the models to determine factors that contribute the pedestrian crashes. The paper concludes that pedestrian crashes often happen in the evenings, with low lighting and visibility levels, independent of the presence of sidewalks.

1.2 Related Literature

The study of vehicle-pedestrian crashes has been examined by several researchers looking at different aspects of the problem. Several statistical methodologies have been used to model pedestrian crashes including mixed logit, logistic regression, ordered probit, and binary logistic regression (*4,5,6,7,8,9,10,11*).

Studies have concluded that increases in speed lead to more severe crashes while increases in lanes and width of lanes tended to decrease the number of crashes (*12,13,14*). The urban environment has been studied and determined that land-use and transit availability have an influence on pedestrian crashes, typically negatively as walking friendly development and increased transit tend to have higher instances of pedestrian crashes, however it is often assumed that these higher numbers are actually lower on a comparative rate bases to the exposure of pedestrian and the number of

people choosing to walk (*15,16,17,18,19,20*). Other factors such as traffic operations have been shown to decrease crashes (*21,22*).

Studies by McMahon and colleagues have concluded the pedestrian crashes tend to be higher in locations where sidewalks do not exist versus locations where sidewalks are present (*23,24*). Another paper by Retting et al. concluded that sidewalks can reduce the risk of pedestrian crashes in residential areas (*25*). With regard to residential areas, several studies indicated that traffic calming devices, intended to reduce the speed of vehicles, also can reduce the number of pedestrian crashes because many were caused by children who often do not accurately gauge speed of vehicles and tend to cross mid-block (*26,27,28,29*). Pedestrian visibility was often cited as an issue in crashes and the installation of lights for nighttime pedestrians was presented as means to improve safety (*30,31*).

This study performs a binary logistic statistical analysis to test the impact of the presence of sidewalks on pedestrian crash severity, which has not been covered in the related literature on pedestrian crashes.

2. Methodology

To analyze the differences in crashes between those that occur with a sidewalk present and those that occur without a sidewalk present, a statistical model will be used to analyze the data. The statistical modeling tool used in this study is Binary Logistic Regression using IBM SPSS Statistics 24.

2.1 Logistic Regression

The goal of using binary logistic regression is similar to any type of modeling analysis, to find the best fit and the most parsimonious model. The distinguishing characteristic of the logistic regression model from a linear regression model is the response variable. In the logistic regression model, the response variable is binary or dichotomous (*32*). The difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression analysis.

2.2 Binary Logistic Regression

Binary logistic Regression estimates the probability that a characteristic is present (e.g. estimate probability of "success") given the values of explanatory variables (*32*).

The definitions of the variables Y and X are as follows:

Let Y, for any subset *i*, be a binary response variable such that $Y_i = 1$ if the trait is present in observation and $Y_i = 0$ if the trait is not present in observation.

Let $X = (X_1, X_2, ..., X_k)$ be a set of explanatory variables which can be discrete, continuous, or a combination. x_i is the observed value of the explanatory variables for observation.

For our analysis, the response variable will be $Y_i = 1$ when aa crash with a certain severity is observed and $Y_i = 0$ if the alternate severity is recorded. The explanatory variables $X_1, X_2, ..., X_k$ will be collected from the crash analysis database as an attempt to define the dependent variable.

Setting up these variables gives us the model (*33*):

$$\pi i=\Pr(Yi=1|Xi=xi)=\exp(\beta 0+\beta 1xi)1+\exp(\beta 0+\beta 1xi)\pi i=\Pr(Yi=1|Xi=xi)=\exp(\beta 0+\beta 1xi)1+\exp(\beta 0+\beta 1xi) \quad (1)$$

or,

$$\text{logit}(\pi i)=\log(\pi i1-\pi i)=\beta 0+\beta 1xi=\beta 0+\beta 1xi1+\ldots+\beta kxiklogit(\pi i)=\log(\pi i1-\pi i)=\beta 0+\beta 1xi=\beta 0+\beta 1xi1+\ldots+\beta kxik \quad (2)$$

Several assumptions must be made for the model to be correct. Firstly, the data set Y_i must be independently distributed, that is, the cases are independent of one another. In addition the distribution of Y is $Bin(n_i, \pi_i)$, i.e., binary logistic regression model assumes binomial distribution of the response (*32*). The dependent variable doesn't need to be normally distributed, but it typically assumes a distribution from an exponential family. The data set does not assume a linear relationship between the dependent and independent variables, but it does assume linear relations between the logit and response variables; $logit(\pi) = \beta_0 + \beta X$ (*33*). Homogeneity of variance doesn't need to be satisfied. Errors need to be independent but not normally distributed. Due to the fact that maximum likelihood estimation is used rather than ordinary least squares to estimate its parameters, the model relies on large-sample approximations.

To determine the goodness of fit for the model, various statistics must be considered, namely the chi-square, deviance G^2 and likelihood ratio test and statistic, ΔG^2 and the Hosmer-Lemeshow test and statistic. For estimating the parameters, the *maximum likelihood estimator* (MLE) for (β_0, β_1) is obtained by finding $(\beta\hat{}0,\beta\hat{}1)(\beta\hat{}0,\beta\hat{}1)$ that maximizes (*33*):

$$L(\beta 0,\beta 1)=\prod i=1N\pi yii(1-\pi i)ni-yi=\prod i=1Nexp\{yi(\beta 0+\beta 1xi)\}1+\exp(\beta 0+\beta 1xi) \quad (3)$$

3. Results

3.1 Data Preparation

The data used in this analysis were extracted from the Critical Analysis Reporting Environment (CARE) maintained by the Center for Advance Public Safety at the University of Alabama. Pedestrian crash data was obtained for Huntsville and Montgomery, two cities similar in size, in Alabama. Both cities are around 200,000 in population and sidewalks availability is limited to selected locations throughout the cities, such that there are several areas without and without sidewalks.

To perform the analysis and generate a sufficient amount of data, pedestrian crash data from both cities were aggregated and organized into two datasets, "Sidewalk Present" and "No Sidewalk Present". There were two levels in the response variable to examine the severity of the pedestrian crash, severe indicating that a fatality or incapacitating injury occurred and not severe indicating a minor injury or no injury occurred.

3.1.1 Sidewalk Present

For data analysis purposes, any incident that occurred within 20 meters of a sidewalk was defined to be a sidewalk present crash. An assumption was made that the presence of a sidewalk implied that the pedestrian was using the sidewalk correctly as there is no mechanism to be certain that the pedestrian was not walking in the roadway near a sidewalk. The total number of crashes in this group from both cities is 149.

3.1.2 No Sidewalk Present

All pedestrian crash data that was not included in the Sidewalk Present group was analyzed as No Sidewalk Present crashes. Total number of crashes in this group from both cities is 120. From an exposure metric, the number of crashes that occurred without the presence of a sidewalk is interesting because these areas would be ones where pedestrian traffic would not typically be expected as there are no sidewalks to encourage walking trips. In addition, areas without sidewalks experience difference in roadway lighting, shoulders availability, edge of pavement maintenance and ditch placement.

3.2 Contributing Circumstances

When examining the pedestrian crashes, there are a number of different elements that can contribute the crash, as recorded by the officer completing the crash report. The crash could have been the fault of the driver or the pedestrian. For the analysis, the primary contributing circumstance has been divided into two groups, pedestrian at-fault and driver at-fault as shown in Table 1.

Table 1. Primary Contributing Circumstance for Pedestrian Crashes

Driver's fault	**Pedestrian's fault**
Driving Under the Influence	Improper Crossing
Aggressive Operation	Lying or Sitting in the roadway
Failed to Yield Right of Way	Pedestrian Under the Influence
Not Visible	
Other - No Improper Movement	
Swerved to Avoid Vehicle	
Wrong Side of Road	
Failed to Yield the Right of Way	
Failure to Obey Sign	
Followed too Closely	
Misjudge Stopping Distance	
Traveling Wrong Way	
Unseen Object/Person	
Vision Obstructed	

3.3 Model Development

The statistical analysis process requires a number of steps to perform. Each step includes a selection of variable and provides summary statistics to evaluate the model. The results from the step methodology can be interpreted using the following matrix shown in Table 2.

Table 2. Step Model Selection Matrix

	Predicted	
Observed	Not Severe	Severe
Not Severe	Model is Predicting Accurately	Acceptable
Severe	Not Acceptable	Model is Predicting Accurately

The horizontal axis is the predicted values from the analysis and the vertical axis compares the results to the observed condition. The node "Severe x Severe" and "Not Severe x Not Severe" are cases in which the model is accurately predicting the observed case. When the model is predicting a severe case and the observation is also a severe case, it is concluded that the model is predicting accurately. When the prediction suggests a severe case but the observation is not severe ("Severe x Not Severe") no issue is raised because this creates a more conservative prediction and builds in a factor of safety. However, the case that is predicted to be not severe but is observed as severe ("Not Severe x Severe") is underestimating the safety of the section. Therefore the combination which produces the least liberal (or most conservative) case is to be chosen as it has the highest factor of safety.

The other value that was involved in the model development task was the overall percentage correct. If multiple steps have the same level of safety, i.e. the lowest amount of underestimated cases, the overall model accuracy percentage is compared to select the most appropriate combination of variables.

3.3.1 Sidewalk Present

Table 3 shows the severe and not severe cases in each step using different variables for locations where there is a sidewalk present. The observed and predicted values are generated and the overall percentage correct shows how accurately the model was predicting the observed condition. The steps are various iterations of combinations of applicable variables in order to determine the best suited combination to most accurately predict the observed case.

Table 3. Best Step Option Model, Sidewalk Present

			Predicted		
			Crash Severity		
	Observed		Not Severe	Severe	Percentage Correct
Step 6	Crash Severity	Not Severe	89	13	87.3
		Severe	30	17	36.2
	Overall Percentage				**71.1**

Using both values identified and Table 3, the model from step 6 was selected for use in this particular analysis. The specific variables and results of the model obtained from the software are shown in Table 4.

Table 4. Results from SPSS, Sidewalk Present

			Estimate	Standard Error	Odds ratio (OR)	95% C.I.	
Time	10:00 AM	Reference Category				Lower	Upper
Time(1)	4:00 PM		-0.017	0.661	0.983	0.269	3.588
Time(2)	7:00 AM		-1.101	0.567	0.333	0.109	1.01
Time(3)	7:00 PM		1.135	0.846	3.11	0.592	16.339
Lighting Condition	Darkness	Reference Category					
Lighting Condition(1)	Daylight		0.817	0.585	2.264	0.72	7.125
Weather Condition	Clear	Reference Category					
Weather Condition(1)	Rain		1.044	0.732	2.84	0.676	11.933
Causal Unit Age	17 to 24	Reference Category					
Causal Unit Age(1)	25 to 54		1.163	0.941	3.199	0.506	20.228
Causal Unit Age(2)	55 to 74		0.421	0.621	1.524	0.451	5.145
Causal Unit Age(3)	CU is No		0.408	0.849	1.503	0.285	7.938
Not at fault Age	0 to 15	Reference Category					
Not at fault Age(1)	16 to 25		0.132	1.53	1.141	0.057	22.881
Not at fault Age(2)	26 to 64		-1.388	1.608	0.25	0.011	5.841
Not at fault Age(3)	65 Years +		-0.389	1.502	0.678	0.036	12.859
Not at fault Age(4)	N/A		0.542	1.697	1.719	0.062	47.84
Causal Unit Gender	Female	Reference Category					
Causal Unit Gender(1)	Male		1.209	1.714	3.351	0.117	96.326
Causal Unit Gender(2)	Unknown		1.861	1.657	6.428	0.25	165.382
Roadway Curvature	Curve	Reference Category					
Roadway Curvature(1)	Other		1.955	1.708	7.067	0.249	200.829
Roadway Curvature(2)	Straight		0.154	0.573	1.166	0.379	3.588
Constant			-3.304	1.228	0.037		

From the data presented in Table 4, the odds ratios show the odds that a severe crash will occur in the variable data set compared to a reference category. For example, in the light condition, the odds ratio says that a severe incident is approximately 2.26 times more likely to occur in darkness than in daylight. This result makes sense because of the difficulty in drivers seeing individuals walking the evening and during dark periods. This result coincides with the time of day odds ratio that shows a much higher likelihood of being in a severe crash after 7:00 PM in the evening. A similar odds

ratio is developed for clear versus rainy weather condition, indicating a much higher likelihood that individuals will be walking when the weather is nice and therefore the potential exposure is greater for individuals to be in a severe crash. For the casual age groups, 25 to 54 has an OR of 3.2, but when looking at the reference category and the other variables, the overall number of drivers in Huntsville from this population are likely influencing the number of severe crashes with pedestrians as the reference category is relatively small, 17 to 24, and the other groups have a larger number of potential drivers than the reference group and tend to drive more miles. For the individuals likely to be involved in a crash as a pedestrian, the most likely age range is 16 to 25. This is also logical as these individuals often take greater risks.

3.3.2 No Sidewalk Present

Table 5 shows the severe and not severe cases in each step using different variables for locations where there is not a sidewalk present. The observed and predicted values are generated and the overall percentage correct shows how accurately the model was predicting the observed condition. The steps are various iterations of combinations of applicable variables in order to determine the best suited combination to most accurately predict the observed case.

Table 5. Best Step Option Model, No Sidewalk Present

			Predicted		
			Crash Severity		Percentage Correct
	Observed		Not Severe	Severe	
Step 3	Crash Severity	Not Severe	46	5	90.2
		Severe	8	25	75.8
	Overall Percentage				**84.5**

Using both values identified and Table 5, the model from step 3 was selected for use in this particular analysis. The specific variables and results of the model obtained from the software are shown in Table 6.

Table 6. Results from SPSS, No Sidewalk Present

			Estimate	**Standard Error**	**Odds Ratio (OR)**	**95% C.I.**	
						Lower	Upper
Day of Week(1)	Weekend		-1.478	.155	.228	.030	1.746
Time	10:00 AM	Reference Category		.541			
Time(1)	4:00 PM		-38.829	.997	.000	.000	.
Time(2)	7:00 AM		-1.385	.158	.250	.037	1.710
Time(3)	7:00 PM		-.483	.770	.617	.024	15.616
Lighting(1)	Daylight		.172	.874	1.188	.141	9.989
Weather(1)	Rain		-22.306	.999	.000	.000	.
Locale	Open Country	Reference Category		.334			
Locale(1)	Residential		.774	.588	2.168	.131	35.785
Locale(2)	Shopping		-1.129	.221	.323	.053	1.970
CUPedMan	Entering	Reference Category		.690			
CUPedMan(1)	N/A		.782	.389	2.186	.370	12.931
CUPedMan(2)	Walking		20.176	.998	578642084	.000	.

V2 Age	15 to 34	Reference Category		1.000			
V2 Age(1)	45 to 64		39.677	.998	1704627395	.000	.
V2 Age(2)	65 Years +		-38.611	.998	.000	.000	.
CUGen	Female	Reference Category		.819			
CUGen(1)	Male		19.429	.998	274131915	.000	.
CUGen(2)	Unknown		18.928	.998	166089394	.000	.
CUCurveGra	Curve	Reference Category		.810			
CUCurveGra(1)	Other		-80.497	.997	.000	.000	.
CUCurveGra(2)	Straight		.702	.516	2.018	.242	16.833
Constant			4.446	1.000	85.251		

From the data presented in Table 6, the odds ratio show that the odds a severe crash will occur in the variable data set compared to the reference category. For example, weekends are more likely to have a pedestrian crash when no sidewalk is present. This could indicate that more individuals are walking on roadways without sidewalks during the weekends. In addition, the data show that during even hours, after 7:00 PM, during non-daylight hours when it is not raining, the likelihood of being involved in a severe pedestrian crash are higher. These conclusions make sense conceptually, as the combination of darkness, nighttime and weekend walking without sidewalks all tend to lead to higher severity pedestrian crashes.

The additional factor of cause, pedestrian under the influence, is a contributing circumstance to increase these crashes as the presents of alcohol or drugs impairs judgement and can lead pedestrians to attempt to cross when there is not a sufficient gap to allow a pedestrian to cross the street or encourage safe walking along a roadway. Finally, residential locations where the roadways are curved tended to lead to higher crash severity for pedestrians. Again, this is logical as most individuals walking at night would be near their residence and the curvature of roadway would obscure the vision of the driver to reduce the reaction time to avoid the crash.

4. Discussion

4.1 Comparison

One difference between the factors for when sidewalks are present versus when sidewalk are not present is that the sidewalk model includes the driver while the no sidewalk present has variable related to the pedestrian. This indicates that the models are assigning different causal units based on the presence of the infrastructure. When sidewalks are present, the crashes are caused by the driver, or at least attributed to the driver by the reporting officer's opinion. Alternatively, when sidewalks are not present, the pedestrian is reported to be responsible for the crash at a much higher and statistically significant level.

4.2 Conclusions

This paper examined pedestrian crash characteristics for severe versus not severe crashes for situations when a sidewalk is present and those when a sidewalk is not present. In both instances, higher severity pedestrian crashes tended to occur in the evening hours, during periods of darkness. This conclusion is important because driver education can be introduced to help expose this issue make drivers aware that pedestrians are out walking during evening hours, and not to assume the because of the hour that pedestrians will not be present along the roadways.

In both sidewalk present and no sidewalk present scenarios, males tend to have a higher likelihood of being in a severe pedestrian crash. This may be attributed to the comfort level of males walking during the evening and darkness hours or may be a reflection of the risk taking attitudes, especially when the presence of alcohol or drugs might be a factor. Additionally, walking on curved roadways was seen in both instances to increase severity as the sight distance is limited. Interesting to note, that the crashes tend to be more severe in residential neighborhood when sidewalks are not present during the weekends; indicating that individuals might be more likely to feel comfortable walking without a sidewalk in residential locations than in commercial areas.

Overall, the comparison indicated that the presence of sidewalks does not lead to an extreme difference between the factors that influence the severity of pedestrian crashes in these two case study cities. Generally, pedestrians are more

likely to be involved in a severe crash when walking during evening hours when the weather is good and visibility is low due to lighting conditions.

References

Al-Ghamdi, Ali, S. (2002). "Using Logistic Regression to Estimate the Influence of Accident Factors on Accident Severity." *Accident Analysis & Prevention*, *34*(6), 729-41. Web.

CARE data from Alabama Crash System. Database available from the Center for Advance Public Safety at the University of Alabama. Http://www.caps.ua.edu.

Clifton, K., & Kreamer-Fults, K. (2007). An Examination of the Environmental Attributes Associated with Pedestrian-vehicular Crashes near Public Schools. Accident Analysis & Prevention, 708–15.

Data on Binary Logistic Regression. Access at https://onlinecourses.science.psu.edu/stat504/node/150

Dumbaugh, E., & Li, W. (2011). Designing for the Safety of Pedestrians, Cyclists, and Motorists in the Built Environment.' *Journal of the American Planning Association, 77*, 69–88.

Embry, D. D., & Malfetti, J. M. (1981). Stay Out of the Street! Reducing the Risk of Pedestrian Accidents to Preschool Children Through Parent Training and Symbolic Modeling. Falls Church, Va: AAA Foundation for Traffic Safety; Safe Playing Project report 2.

Haleem, K., Alluri, P., & Gan, A. (2015). Analyzing pedestrian crash injury severity at signalized and non-signalized locations. *Accident Analysis & Prevention, 81*, 14–23.

Islam, S., & Jones,. S. L. (2014) Pedestrian at-fault crashes on rural and urban roadways in Alabama. *Accident Analysis & Prevention, 72*, 267–276.

Jacobsen, P. L. (2003). Safety in Numbers: More Walkers and Bicyclists, Safer Walking and Bicycling. *Injury Prevention, 9*, 205–209.

Kim, J., Ulfarsson, G. F., Shankar, V. N., & Mannering, F. L. (2010). "A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. *Accident Analysis & Prevention, 42*(6), 1751–1758.

Kim, J., Ulfarsson, G. F., Shankar, V. N., & Kim, S. (2008). Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis. *Accident Analysis & Prevention, 40*(5), 1695–1702.

Kim, K., Brunner, I. M., & Yamashita, E. (2008). Modeling fault among accident—Involved pedestrians and motorists in Hawaii. *Accident Analysis & Prevention, 40*(6), 2043–2049.

Kitali, A. E., & et al. (2017). Evaluating Aging Pedestrian Crash Severity Using Bayesian Complementary Log-Log Model for Improved Prediction Accuracy. Proceeding of the Transportation Research Board Annual Meeting. No. 17-06386.

Kraus, J. F., Hooten, E. G., Brown, K. A., Peek-Asa, C., Heye, C., & McArthur, D. (1996). Child pedestrian and bicyclist injuries: results of community surveillance and a case-study control. *Inj Prev. 2*, 212–218. Crossref, Medline.

Kwayu, K. M., Kwigizile, V., & Oh, J. (2016). Investigating the Correlation Between Factors Contributing to Pedestrian Involved Crashes and Their Impact on Crash Severity. Transportation Research Board Annual Meeting, Washington, D.C.

Lee, C., & Abdel-Aty,. M. (2005). Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. *Accident Analysis & Prevention, 37*(4), 775–786.

McMahon, P. J. (2002). An analysis of factors contributing to" walking along roadway" crashes: research study and guidelines for sidewalks and walkways. Vol. 1. DIANE Publishing.

McMahon, P., & et al. (1999). "Analysis of factors contributing to" walking along roadway" crashes." Transportation Research Record: *Journal of the Transportation Research Board*, 1674, 41-48.

(2014). National Highway Traffic Safety Administration. Traffic Safety Facts, Accessed at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812270

Noland, R. B., & Oh, L. (2004). The Effect of Infrastructure and Demographic Change on Traffic-related Fatalities and Crashes: A Case Study of Illinois County-level Data. *Accident Analysis & Prevention, 36*(4), 525–532.

Pegrum, B. V. (1972). The application of certain traffic management techniques and their effect on road safety. In: Proceedings of the National Road Safety Symposium. Perth, Western Australia: Dept of Shipping and Transport, 277–286.

Phinney, J., Colker, L., & Cosgrove, M. (1985). Literature Review on the Preschool Pedestrian. Washington, DC: US

Dept of Transportation;. DOT publication HS-806-679.

Polus, A., & Katz, A. (1978). An analysis of nighttime pedestrian accidents at specially illuminated crosswalks. Accid Anal Prev.;10:223–228. Crossref.

Retting, R. A., Susan, A., F., & Anne, T. M. (2003). "A review of evidence-based traffic engineering measures designed to reduce pedestrian–motor vehicle crashes." *American journal of public health, 93*(9), 1456-1463.

Rosén, E., & Ulrich, S. (2009). Pedestrian fatality risk as a function of car impact speed. *Accident Analysis & Prevention, 42*, 536-542.

Rothman, L., William, H. A., Camden, A., & Macarthur, C. (2012). Pedestrian crossing location influences injury severity in urban areas. Injury prevention.

Sharma, A., & et al. (2017). *Leading Pedestrian Interval Implementation as a Marginal Costs and Benefits Problem*. Proceeding of the Transportation Research Board Annual Meeting. No. 17-05116.

Stoker, P., Garfinkel-Castro, A., Khayesi, M., Odero, W., Mwangi, M. N., Peden, M., & Ewing, R. (2015). Pedestrian Safety and the Built Environment: A Review of Risk Factors. *Journal of Planning Literature, 30*(4), 377-392.

Welch, E. A. (2017) Identifying factors explaining pedestrian crash severity: a study of Austin, Texas. Diss.

Yu, Chia-Yuan. (2015). Built Environmental Designs in Promoting Pedestrian Safety. *Journal of Sustainability, 1*(7), 9444-9460.

Zajac, S., & John, I. (2003). Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural Connecticut. *Accident Analysis and Prevention, 35*(3), 369-379.

Zeedyk, M. S., Wallace, L., & Spray, L. (2002). Stop, look, listen and think? What young children really do when crossing the road. *Accid Anal Prev., 34*, 43–50. Crossref, Medline.

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Stress-Strength Reliability Model with The Exponentiated Weibull Distribution: Inferences and Applications

Fathy H. Eissa[1,2]

[1] College of Science and Arts- Rabigh, King Abdulaziz University, Saudi Arabia

[2] Department of Mathematics, Faculty of Science, Damanhour University, Damanhour, Egypt

Correspondence: Fathy H. Eissa, Department of Mathematics, Faculty of Science, Damanhour University, Damanhour, Egypt.

Received: May 2, 2018 Accepted: May 18, 2018 Online Published: June 26, 2018

doi:10.5539/ijsp.v7n4p78 URL: https://doi.org/10.5539/ijsp.v7n4p78

Abstract

In this paper, we deal with the estimation of the reliability $R = P(Y < X)$ where X, a unit strength, and Y, a unit stress, are independent exponentiated Weibull random variables. The maximum likelihood and Bayesian methods are used to make inference about R. We obtain the Baysian estimator using Lindely's procedure under squared error loss and LINEX loss functions with gamma prior for the unknown model parameters. The asymptotic and bootstrap confidence intervals are obtained as well as the credible interval for R is constructed in view of the empirical Bayesian procedure. For illustrative purposes, analysis of real data sets is presented. Mont Carlo simulations are carried out to compare the performances of the different estimators.

Keywords: maximum likelihood estimation, stress-strength model, Lindely's approximation, asymptotic confidence interval, bootstrap intervals, credible interval

1. Introduction

We consider the inference on the reliability $R = P(Y < X)$ of a system where X, a unit strength, and Y, a unit stress, are independent exponentiated Weibull random variables. This function means that R is the probability that a system is strong enough to overcome the stress imposed on it. The reliability parameter R is a measure of a system performance. Birnbaum (1956) who was introduced the main idea of this area of research. The stress-strength model has wide applications in several fields. For example, in engineering, X can represent the strength of a system structure and Y represents the stress due to environmental conditions imposed on it. Information about the mechanical reliability of system design can be obtained prior the production through stress- strength model. This information can decrease the costs of production. Other example, in biology, R can be a measure of the difference between two populations and has applications in many areas. When X is a treatment group and Y represents a control group, R refers to a measure of the treatment effects. For details, see Hauk et al. (2000), Reiser (2000) and Wellek (1993). Due to the practical importance, the estimation of R has attracted the attention of several authors who considered several distributions such as exponential, normal, Weibull, generalized exponential etc.. Among of other works deal with inferences about R: Mahdizadeh (2018), Sarhan et al. (2015), Rao et al. (2016), Jovanovic and Rajic (2014), Raqab et al. (2008), Weerahndi and Johnson (1992), Constantine et al. (1986), Rezaei et al. (2010). Our aim in this research is to focus on inferences for $R = P(Y < X)$ when X and Y are two independent but not identical distributed random variables with the exponentiated Weibull (EW) distribution. We use several estimation methods: classical and Bayesian for point estimation and asymptotic confidence, bootstrap confidence intervals and credible interval for interval estimation. The performances of Bayes and non-Bayes methods are compared by analysis of real data sets and Mont Carlo simulations through computed the mean square error of different estimators and average lengths and coverage probability of different estimating intervals. The exponentiated Weibull random variable has a cumulative distribution function

$$F(x) = (1 - e^{-x^{\alpha}})^{\theta} \tag{1}$$

and the corresponding probability density function (pdf)

$$f(x) = \alpha\theta x^{\alpha-1} e^{-x^{\alpha}} (1 - e^{-x^{\alpha}})^{\theta-1}, x > 0, \quad \alpha \quad and \quad \theta > 0. \tag{2}$$

Here α and θ are shape parameters. We use the abbreviation $EW(\alpha, \theta)$ to denote the exponentiated Weibull distribution with density cited above. This distribution has been introduced by Mudholkar and Srivastava (1993). The EW family includes many important distributions. For examples, for $\theta = 1$, it represents Weibull distribution, for $\alpha = 1$, it represents the exponentiated exponential distribution. For $\alpha = 2$, it represents the one-parameter Burr type-X distribution as well as a generalized Rayleigh distribution. Furthermore, The EW distribution has a convenient structure of its distribution

function that can be used quite adequately and effectively in analyzing several lifetime data. The article is organized as follows: In Section 2, we consider the maximum likelihood estimation. In Section 3, we derive different confidence intervals estimation for R. Section 4 proposes Bayesian approximation technique to get the Bayesian estimation for R. Section 5, adopts empirical Bayesian procedure to obtain a credible interval estimation for R. Analysis of real data sets is given in Section 6. In Section 7 simulation study is carried out, and Section 8 concludes the paper.

Now, we assume that X follows $EW(\alpha_1, \theta_1)$ and Y follows $EW(\alpha_2, \theta_2)$. Our interesting value is the reliability parameter R defined by

$$R = P(Y < X) = E_X(F_Y(x)).$$

Using this form with equation (2), we get

$$R = \alpha_1\theta_1 \int_0^\infty x^{\alpha_1-1}e^{-x^{\alpha_1}}(1 - e^{-x^{\alpha_1}})^{(\theta_1-1)}(1 - e^{-x^{\alpha_2}})^{\theta_2-1)}dx.$$

Applying the series expansion $(1 - z)^a = \sum_{i=0}^{\infty} \frac{(-1)^i\Gamma(a+1)z^i}{\Gamma(a+1-i)i!}$,on the last two terms of the integrand with some mathematical manipulations, we get, finally, the form of R as

$$R = \theta_1\Gamma(\theta_1)\Gamma(\theta_2 + 1)\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\frac{(-1)^{i+j+k}(i + 1)^{-\alpha_1(k\frac{\alpha_2}{\alpha_1}+1)}}{i!\,j!k!\Gamma(\theta_1 - i)\Gamma(\theta_2 + 1 - j)}\Gamma(k\frac{\alpha_2}{\alpha_1} + 1). \tag{3}$$

Alternatively, we assume that $\alpha_1 = \alpha_2 = \alpha$ and then the reliability parameter R can be obtained as

$$R = \frac{\theta_1}{(\theta_1 + \theta_2)}. \tag{4}$$

The assumption of this form may be associated with many practical situations. If $\theta_1 = \theta_2$, $R = 0.5$, that is X and Y are independent and identically distributed and there is an equal chance that the strength is greater than stress. When θ_1 and θ_2 are estimated the value of R is simply estimated using equation (4). We remark that equation (4) does not contain α but θ_1 and θ_2 are functions of α and hence R depends on α. However, if α is (estimated) already known, the estimators of θ_1 and θ_2 are obtained and hence so does the estimator of R.

2. Maximum Likelihood Estimation

Suppose $\underline{x} = \{x_1, x_2, \ldots, x_{n_1}\}$ and $\underline{y} = \{y_1, y_2, \ldots, y_{n_2}\}$ be two random samples taken from $EW(\alpha, \theta_1)$ and $EW(\alpha, \theta_2)$, respectively. The observed value x_i represents the strength of $i - th$ component and observed value y_i represents the stress acting on it. Based on these observed samples, the likelihood function of α, θ_1 and θ_2 is

$$L(\underline{x}, \underline{y}|\alpha, \theta_1, \theta_2) \propto \alpha^{n_1+n_2}\theta_1^{n_1}\theta_2^{n_2}e^{-(T_1+T_2)} \tag{5}$$

The log-likelihood function, l, is

$$l \propto nln\alpha + n_1ln\theta_1 + n_2ln\theta_2 - T_1 - T_2 \tag{6}$$

where

$$T_1 = \sum_{i=1}^{n_1}[x_i^\alpha - (\alpha - 1)lnx_i - (\theta_1 - 1)lnu_i],$$

$$T_2 = \sum_{i=1}^{n_2}[y_i^\alpha - (\alpha - 1)lny_i - (\theta_1 - 1)lnv_i],$$

$$u_i = 1 - e^{-x_i^\alpha}, \quad v_i = 1 - e^{-y_i^\alpha} \quad and \quad n = n_1 + n_2$$

and the estimating equations can be obtained as

$$\frac{n}{\alpha} - \frac{1}{\alpha}(p_1 - q_1) + \theta_1p_2 + \theta_2q_2 - (p_2 + q_2) = 0, \tag{7}$$

$$\frac{n_1}{\theta_1} + \sum_{i=1}^{n_1} lnu_i = 0, \tag{8}$$

$$\frac{n_2}{\theta_2} + \sum_{i=1}^{n_1} lnv_i = 0 \tag{9}$$

where

$$p_1 = p_1(\alpha) = \sum_{i=1}^{n_1}(x_i^\alpha - 1)lnx_i^\alpha, \quad p_2 = p_2(\alpha) = \sum_{i=1}^{n_1} u_i^{-1} x_i^\alpha e^{-x_i^\alpha} lnx_i,$$

$$q_1 = q_1(\alpha) = \sum_{i=1}^{n_2}(y_i^\alpha - 1)lny_i^\alpha, \quad q_2 = q_2(\alpha) = \sum_{i=1}^{n_2} v_i^{-1} y_i^\alpha e^{-y_i^\alpha} lny_i.$$

From equations (8) and (9), we obtain the ML estimators:

$$\hat{\theta}_1(\hat{\alpha}) = n_1 / \sum_{i=1}^{n_1} lnu_i^{-1}, \quad \hat{\theta}_2(\hat{\alpha}) = n_2 / \sum_{i=1}^{n_2} lnv_i^{-1} \tag{10}$$

where $\hat{\alpha}$ can be obtained as the solution of the nonlinear equation

$$\frac{1}{\alpha}(n - p_1 - q_1) = n_1 p_2 (\sum_{i=1}^{n_1} lnu_i)^{-1} + n_2 q_2 (\sum_{i=1}^{n_2} lnv_i)^{-1} + (p_2 + q_2)$$

that can be rewritten in the form

$$g(\alpha) = \alpha \tag{11}$$

where $g(\alpha) = \frac{n - p_1 - q_1}{p_2[1 + n_1(\sum_{i=1}^{n_1} lnu_i)^{-1}] + q_2[1 + n_2(\sum_{i=1}^{n_2} lnv_i)^{-1}]}$.

The ML estimator, $\hat{\alpha}$, of α can be obtained from equation (11) by using a simple iterative technique as $g(\alpha_{(i)}) = \alpha_{(i+1)}$, where $\alpha_{(i)}$ is the $j-th$ iterate of $\hat{\alpha}$. The iterations should be finished when the absolute value of $(\alpha_{(i)} - \alpha_{(i+1)})$ is sufficiently small. Once $\hat{\alpha}$ is obtained, we get θ_1 and θ_2 using equations (10) and hence the MLE of R is given by

$$\hat{R}_M = \hat{\theta}_1 / (\hat{\theta}_1 + \hat{\theta}_2) \tag{12}$$

on the basis of the invariance property of the MLE.

3. Confidence Intervals

Although $\hat{R_M}$ can be obtained in explicit form, it is difficult to obtain the exact distribution of it. Hence, we mainly depend on the asymptotic distribution of $\hat{R_M}$ to construct an asymptotic confidence interval (ACI) of R. We also consider two different parametric bootstrap confidence intervals.

3.1 Asymptotic Confidence Interval

From the asymptotic distribution of $\hat{\gamma} = (\hat{\theta_1}, \hat{\theta_2}, \hat{\alpha})'$ we derive the asymptotic distribution of $\hat{R_M}$ and hence we obtain the ACI of R. The MLE of $\gamma = (\theta_1, \theta_2, \alpha)'$ is asymptotically normal with mean of true γ and variance-covariance matrix $I^{-1}(\gamma) = (a_{ij}(\gamma))^{-1}$ where $I^{-1}(\gamma)$ is the inverse of the Fisher information matrix $I(\gamma) = -E(\frac{\partial^2 l}{\partial\gamma_i \partial\gamma_j})$, $i, j = 1, 2, 3$. $I(\gamma)$ is consistently estimated by $I(\hat{\gamma})$ where $\hat{\gamma}$ is the MLE of γ. The variance-covariance matrix can be written in terms of its elements as the inverse of the matrix

$$(a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

where the elements a_{ij} for $i, j = 1, 2, 3$ are the negative of second derivatives of the log-likelihood function given by equation (6); That is,

$$a_{11} = \frac{n_1}{\theta_1^2}, \quad a_{22} = \frac{n_2}{\theta_2^2},$$

$$a_{33} = \frac{1}{\alpha}(g_1 + g_2) - \frac{1}{\alpha^2}(p_1 + q_1 - n) - (\theta_1 - 1)h_1 - (\theta_2 - 1)h_2,$$

$$a_{12} = a_{21} = 0, \quad a_{13} = a_{31} = -p_2, \quad a_{23} = a_{32} = -q_2. \tag{13}$$

where p_1, q_1, p_2, q_2 are defined in equation (7),

$$g_1 = g_1(\alpha) = \sum_{i=i}^{n_1}(x_i^\alpha lnx_i + x_i^\alpha - 1)lnx_i,$$

$$g_2 = g_2(\alpha) = \sum_{i=i}^{n_2}(y_i^{\alpha} lny_i + x_i^{\alpha} - 1)lny_i,$$

$$h_1 = h_1(\alpha) = \sum_{i=i}^{n_1}(1 - \varphi_i - x_i^{\alpha})\varphi_i(lnx_i)^2,$$

$$h_2 = h_2(\alpha) = \sum_{i=i}^{n_2}(1 - \psi_i - y_i^{\alpha})\psi_i(lny_i)^2,$$

$$\varphi_i = u_i^{-1}x_i^{\alpha}e^{-x_i^{\alpha}} \quad and \quad \psi_i = v_i^{-1}y_i^{\alpha}e^{-y_i^{\alpha}}.$$

The MLE is $\hat{R}_M = \hat{\theta_1}/(\hat{\theta_1} + \hat{\theta_2})$ as given by equation (12), is asymptotically normally distributed with mean R and variance $\sigma_R^2 = \sum_{=1}^{2}\sum_{j=1}^{2}\frac{\partial R}{\partial\theta_i}\frac{\partial R}{\partial\theta_j}I_{ij}^{-1}(\gamma)$ (Rao 1973) which is consistently estimated to be

$$\sigma_R^2 = \frac{1}{J(\theta_1 + \theta_2)^4}[(a_{11}a_{33} - a_{13}^2)\theta_1^2 - 2a_{13}a_{23}\theta_1\theta_2 + (a_{22}a_{33} - a_{23}^2)\theta_2^2] \tag{14}$$

where $J = a_{11}a_{22}a_{33} - a_{11}a_{23}^2 - a_{22}a_{13}^2$.
Remembring that all the above values of $var(\hat{R_M}) = \sigma_R^2$ is computed at the MLE of the parameters θ_1, θ_2 and α. Therefore, an asymptotic $100(1-\tau)\%$ confidence interval, ACI, for R can be obtained as

$$[\hat{R}_M + z_{\tau/2}\sigma_R, \hat{R}_M - z_{\tau/2}\sigma_R]. \tag{15}$$

where z_k is the $k-th$ quantile of the standard normal distribution. A better of such confidence interval may be obtained in cases of large sample sizes. For small sample sizes, we adopt the bootstrap confidence interval in the following.

3.2 Bootstrap Confidence Intervals

In this section, we propose the use of the following method to generate parametric bootstrap samples, suggested by Efron and Tibshirani (1998), of R, starting from the given independent random samples $\underline{x}$ and $\underline{y}$ obtained from $EW(\alpha, \theta_1)$ and $EW(\alpha, \theta_2)$, respectively. We employ the percentile bootstrap and Student's t bootstrap confidence intervals for R. The steps of the method to construct the bootstrap confidence interval for R are summarized in the following steps:

Step 1. Given a random sample $\underline{x} = \{x_1, x_2, \ldots, x_{n_1}\}$ and $\underline{y} = \{y_1, y_2, \ldots, y_{n_1}\}$, calculate $\hat{\alpha}, \hat{\theta}_1$ and $\hat{\theta}_2$.
Step 2. Sample with replacement from the original sample using α, θ_1 and θ_2 computed in step 1. Generate a bootstrap sample $\underline{x}^* = \{x_1^*, x_2^*, \ldots, x_{n_1}^*\}$ using $\hat{\alpha}$ and $\hat{\theta}_1$ and similarly generate $\underline{y}^* = \{y_1^*, y_2^*, \ldots, y_{n_1}^*\}$ using $\hat{\alpha}$ and $\hat{\theta}_2$.
Step 3. Calculate the same statistics $\hat{\alpha}^*$, $\hat{\theta_1}^*$ and $\hat{\theta_2}^*$ as in step 1 using the sample found in step 2. Compute the bootstrap estimate of R using equation (12), say $\hat{R}^*$.
Step 4. Repeat steps 2-3, N times, where $N \geq 1000$, and put the bootstrap values $\hat{R}^*$ in ascending order.

(i) Percentile bootstrap ($p-boot$) confidence interval
Define $\hat{R}^{*(p)}$ such that $(\frac{1}{N})\sum_{j=1}^{N} I(\hat{R}_j^* \leq \hat{R}^{*(p)}) = p$ where $\hat{R}^{*(p)}$ is the p percentile of $\{\hat{R}_j^*, j = 1, \ldots, N\}$, $0 < p < 1$ and $I(..)$ is the indicator function.
The $(1-\tau)100\%$ $p-boot$ confidence interval for R is given by

$$[\hat{R}^*(\tau/2), \hat{R}^*(1-\tau/2)]. \tag{16}$$

(ii) Student's t bootstrap ($t-boot$) confidence interval
Consider the sample mean, $\bar{\hat{R}}^* = (1/N)\sum_{j=1}^{N}\hat{R}_j^*$, and sample variance, $Var(\hat{R}^*) = (1/N)\sum_{j=1}^{N}(\hat{R}_j^* - \bar{\hat{R}}^*)^2$ of $\{\hat{R}_j^*, j = 1, \ldots, N\}$. Define statistic $\hat{T}^{*(p)}$ such that $(1/N)\sum_{j=1}^{N} I(\frac{\hat{R}_j^* - \hat{R}_M}{\sqrt{Var(\hat{R}^*)}} \leq \hat{T}^{*(p)}) = p$ where $\hat{T}^{*(p)}$ is the p percentile of $\{\frac{\hat{R}_j^* - \hat{R}_M}{\sqrt{Var(\hat{R}^*)}}, j = i, \ldots, N\}$. The $(1-\tau)100\%$ $t-boot$ confidence interval for R is given by

$$[\hat{R}_M + \hat{T}^{*(\tau/2)}\sqrt{Var(\hat{R}^*)}, \hat{R}_M - \hat{T}^{*(\tau/2)}\sqrt{Var(\hat{R}^*)}]. \tag{17}$$

4. Bayesian Estimation of R

In this section, the Bayes estimates of R are obtained. We assume that the parameters θ_1, θ_2 and α have independent gamma distributions, priori, each with density function $\Pi(\gamma) \propto \gamma^{a-1}e^{-b\gamma}$, $\gamma > 0$, for fixed values of a, $b > 0$ and γ is the vector space $(\theta_1, \theta_2, \alpha)'$. The joint posterior density function of θ_1, θ_2 and α can be obtained as

$$p(\theta_1, \theta_2, \alpha|\underline{x}, \underline{y}) = k^{-1}\alpha^{n+a_0-1}\theta_1^{n_1+a_1-1}\theta_2^{n_2+a_2-1}e^{-\delta_1\theta_1}e^{-\delta_2\theta_2} \times e^{-(b_0+z)\alpha}e^{-c+d+z} \tag{18}$$

where $\delta_1 = \delta_1(\alpha) = b_1 + d_1$, $\delta_2 = \delta_2(\alpha) = b_2 + d_2$,
$z = \sum_{i=1}^{n_1} ln(x_i)^{-1} + \sum_{i=1}^{n_2} ln(y_i)^{-1}$, $c = c(\alpha) = \sum_{i=1}^{n_1} x_i^{\alpha} + \sum_{i=1}^{n_2} y_i^{\alpha}$,
$d_1 = d_1(\alpha) = \sum_{i=1}^{n_1} lnu_i^{-1}$, $d_2 = d_2(\alpha) = \sum_{i=1}^{n_2} lnv_i^{-1}$, $d = d_1 + d_2$, $n = n_1 + n_2$,
u_i and v_i are given in equation (6) and k^{-1} is the normalizing constant. The Byes estimator of R under squared error loss function is given by

$$\hat{R}_B = \int_0^{\infty}\int_0^{\infty}\int_0^{\infty} R(\theta_1, \theta_2)p(\theta_1, \theta_2, \alpha|\underline{x}, \underline{y})d\alpha d\theta_1 d\theta_2. \tag{19}$$

In view of difficulty to evaluate the posterior expectation in equation (19) analytically, we employed Lindely's approximation method to approximate the ratio of integrals in equation (19) and so we can obtain the estimate of R. Depending on the ML estimators for α, θ_1, and θ_2, we use lindely's approximation form expanding about these estimators.
Lindely's approximation:
Lindely (1980) developed an approximate procedure to evaluate the ratio of two integrals such as that of the posterior mean of a function $w(\lambda)$ where

$$E(W(\lambda)|t) = \int w(\lambda)e^{q(\lambda)}d\lambda / \int e^{q(\lambda)}d\lambda \tag{20}$$

where $q(\lambda) = l(\lambda) + \rho(\lambda)$, $l(\lambda)$ is the logarithm of the likelihood function and $\rho(\lambda)$ is the logarithm of the prior density of λ where λ is a vector of parameters, say $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r)$. According to Lndely's approximation, $E(W(\lambda)|\underline{t})$ is approximately estimated by the form

$$E(W(\lambda)|\underline{t}) = [w + (1/2)\sum_i\sum_j(w_{ij} + 2w_i\rho_i)\sigma_{ij} + (1/2)\sum_i\sum_j\sum_k\sum_l l_{ijk}\sigma_{ij}\sigma_{kl}w_l]_{\lambda=\hat{\lambda}} + termsofordern^{-2}orsmaller \tag{21}$$

where $w = w(\lambda)$, $i, j, k, l = 1, 2, 3, \ldots, r$, $w_i = \partial w/\partial\lambda_i$, $w_{ij} = \partial^2 w/\partial\lambda_i\partial\lambda_j$, $l_{ijk} = \partial^3 l/\partial\lambda_i\lambda_j\lambda_k$, $\rho_j = \partial\rho/\partial\lambda_j$, σ_{ij} is the $(i, j)th$ element in the inverse of the matrix $\{-l_{ij}\}$ and $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_r)$ is the MLE of λ, viz, all these quantities are evaluated at the MLE of the parameters. Consider the case of three parameters; that is when $\lambda = (\lambda_1, \lambda_2, \lambda_3)$.The posterior mean from equation (21)) is reduced to

$$\begin{aligned}\hat{w}_B = E(W(\lambda)|\underline{t}) = w + (w_1\delta_1 + w_2\delta_2 + w_3\delta_3 + \delta_4 + \delta_5) + (1/2)[A(w_1\sigma_{11} + w_2\sigma_{12} + w_3\sigma_{13}) + \\ B(w_1\sigma_{21} + w_2\sigma_{22} + w_3\sigma_{23}) + C(w_1\sigma_{31} + w_2\sigma_{32} + w_3\sigma_{33})]\end{aligned} \tag{22}$$

where

$\delta_i = \sum_{j=1}^{3}\rho_j\sigma_{ij}$, $i = 1, 2, 3$,
$\delta_4 = w_{12}\sigma_{12} + w_{13}\sigma_{13} + w_{23}\sigma_{23}$, $\delta_5 = (1/2)(w_{11}\sigma_{11} + w_{22}\sigma_{22} + w_{33}\sigma_{33})$,
$A = \sigma_{11}l_{111} + 2\sigma_{12}l_{121} + 2\sigma_{13}l_{131} + 2\sigma_{23}l_{231} + \sigma_{22}l_{221} + \sigma_{33}l_{331}$,
$B = \sigma_{11}l_{112} + 2\sigma_{12}l_{122} + 2\sigma_{13}l_{132} + 2\sigma_{23}l_{232} + \sigma_{22}l_{222} + \sigma_{33}l_{332}$,
$C = \sigma_{11}l_{113} + 2\sigma_{12}l_{123} + 2\sigma_{13}l_{133} + 2\sigma_{23}l_{233} + \sigma_{22}l_{223} + \sigma_{33}l_{333}$.

In our case, we have $\lambda = (\theta_1, \theta_2, \alpha)$ and $w = w(\theta_1, \theta_2, \alpha) = R$ as given in equation (4). To apply Lidely's form of equation (22), we first obtain the σ_{ij} elements of the inverse of the matrix $\{-l_{ij}\}$, $i, j = 1, 2, 3$. From the log-likelihood function given in equation (5), we can obtain σ_{ij} as follows:

$\sigma_{11} = J^{-1}(a_{22}a_{33} - a_{23}^2)$, $\sigma_{22} = J^{-1}(a_{11}a_{33} - a_{13}^2)$, $\sigma_{12} = J^{-1}(a_{13}a_{32} - a_{12}a_{33}) = \sigma_{21}$, $\sigma_{13} = J^{-1}(a_{12}a_{23} - a_{13}a_{22}) = \sigma_{31}$,
$\sigma_{23} = J^{-1}(a_{13}a_{21} - a_{11}a_{23}) = \sigma_{32}$,

where a_{ij}, $i, j = 1, 2, 3$ are given by equations (13) and J is given in equation (14).

The quantities ρ_j and l_{ijk}, $i = 1, 2, 3$ are obtained as

$\rho_1 = (a_1 - 1)\theta_1^{-1} - b_1$, $\rho_2 = (a_2 - 1)\theta_2^{-1} - b_2$, $\rho_3 = (a_0 - 1)\alpha^{-1} - b_0$,
$l_{111} = 2n_1\theta_1^{-3}$, $l_{222} = 2n_2\theta_2^{-3}$,

$l_{333} = \alpha^{-2}(f_1 + f_2 + g_1 + g_2) - 2\alpha^{-3}(p_1 + q_1 - n) - 3\alpha^{-1}(k_1 + k_2) + (\theta_1 - 1)z_1 + (\theta_2 - 1)z_2$,
$l_{133} = l_{331} = h_1$, $l_{233} = l_{332} = h_2$

where $f_1 = \sum_{i=1}^{n_1}(x_i^\alpha lnx_i^\alpha + x_i^\alpha - 1)lnx_i$, $f_2 = \sum_{i=1}^{n_2}(y_i^\alpha lny_i^\alpha + y_i^\alpha - 1)lny_i$,
$z_1 = \sum_1^{n_1}[2\varphi_i^3 + 3(x_i^\alpha - 1)\varphi_i^2 + (x_i^{2\alpha} - 3x_i^\alpha + 1)\varphi_i](lnx_i)^3$,
$z_2 = \sum_1^{n_2}[2\psi_i^3 + 3(y_i^\alpha - 1)\psi_i^2 + (y_i^{2\alpha} - 3y_i^\alpha + 1)\psi_i](lny_i)^3$,
$k_1 = \sum_1^{n_1} x_i^\alpha(lnx_i)^2$, $k_2 = \sum_1^{n_2} y_i^\alpha(lny_i)^2$,
$h_1, h_2, g_1, g_2, \varphi_i, \psi_i, p_1$ and q_1 are given in equation (13).

Then,

$A = \sigma_{11}l_{111} + \sigma_{33}l_{331}$, $B = \sigma_{22}l_{222} + \sigma_{33}l_{332}$, $C = 2\sigma_{13}l_{133} + 2\sigma_{23}l_{233} + \sigma_{33}l_{333}$, $\delta_1 = J^{-1}A_1$, $\delta_2 = J^{-1}A_2$, $\delta_3 = J^{-1}A_3$

where

$A_1 = (a_{22}a_{33} - a_{23}^2)[(a_1 - 1)\theta_1^{-1} - b_1] + a_{13}a_{32}[(a_2 - 1)\theta_2^{-1} - b_2] + a_{13}a_{22}[(a_0 - 1)\alpha^{-1} - b_0]$,

$A_2 = a_{13}a_{32}[(a_1 - 1)\theta_1^{-1} - b_1] + a_{11}a_{33}[(a_2 - 1)\theta_2^{-1} - b_2] + a_{11}a_{23}[(a_0 - 1)\alpha^{-1} - b_0]$,

$A_3 = -a_{13}a_{22}[(a_1 - 1)\theta_1^{-1} - b_1] - a_{11}a_{23}[(a_2 - 1)\theta_2^{-1} - b_2] + a_{11}a_{22}[(a_0 - 1)\alpha^{-1} - b_0]$.

Moreover, $w_1 = t_1$, $w_1 = t_2$, $w_{11} = t_3$, $w_{22} = t_4$, $w_{12} = t_5$, where

$t_1 = \theta_2(\theta_1 + \theta_2)^{-2}$, $t_2 = -\theta_1(\theta_1 + \theta_2)^{-2}$, $t_3 = -2\theta_2(\theta_1 + \theta_2)^{-3}$, $t_4 = 2\theta_1(\theta_1 + \theta_2)^{-3}$, $t_5 = 2\theta_1(\theta_1 + \theta_2)^{-3} - (\theta_1 + \theta_2)^{-2}$;
$w_3 = w_{33} = w_{13} = w_{23} = 0$.
Also, $\delta_4 = J^{-1}a_{13}a_{32}t_5$, $\delta_5 = 2J^{-1}[(a_{22}a_{33} - a_{23}^2)t_3 + (a_{11}a_{33} - a_{13}^2)t_4]$.
Therefore, The Bayes estimator for R, under squared error loss function and LINEX loss function, using Lindely's approximation can be obtained in what follows.

- Under squared error loss function
The Bayes estimator for R, denoted by $\hat{R}_{BSL}$, under squared error loss function can be evaluated by the form

$$\hat{R}_{BSL} = R + \Phi + \Psi_1 t_1 + \Psi_2 t_2 \tag{23}$$

where $\Phi = (1/2)t_3\sigma_{11} + (1/2)t_4\sigma_{22} + t_5\sigma_{12}$, $\Psi_1 = \delta_1 + (1/2)(A\sigma_{11} + B\sigma_{21} + C\sigma_{31})$,
$\Psi_2 = \delta_2 + (1/2)(A\sigma_{12} + B\sigma_{22} + C\sigma_{32})$, $A = \sigma_{11}l_{111} + \sigma_{33}l_{331}$, $B = \sigma_{22}l_{222} + \sigma_{33}l_{332}$,
$C = 2\sigma_{13}l_{133} + 2\sigma_{23}l_{233} + \sigma_{33}l_{333}$.

All these values are evaluated at the MLEs of θ_1, θ_2 and α.

- Under LINEX loss function

Under LINEX loss function, the Bayes estimator of $w = w(\theta_1, \theta_2, \alpha)$ is given by

$\hat{w}_B = -(1/s)lnE(e^{-sw}|\underline{x}, \underline{y})$, $s \neq 0$.

where
$E(e^{-sw}|\underline{x}, \underline{y}) = \int\int\int_{\theta_1,\theta_2,\alpha} e^{-sw}p(\theta_1, \theta_2, \alpha|\underline{x}, \underline{y})d\theta_1 d\theta_2 d\alpha / \int\int\int_{\theta_1,\theta_2,\alpha} p(\theta_1, \theta_2, \alpha|\underline{x}, \underline{y})d\theta_1 d\theta_2 d\alpha$.
We apply Lindely's approximation on this integral form as were used to evaluate equation (20), to obtain

$$E(e^{-sw}|\underline{x}, \underline{y}) = e^{-sw} + \Phi + \Psi_1 w_1 + \Psi_2 w_2 \tag{24}$$

where $\Phi = (1/2)w_{11}\sigma_{11} + (1/2)w_{22}\sigma_{22} + w_{12}\sigma_{12}$.

The values of w_1, w_2, w_{11}, w_{22} and w_{12} can be obtained as follows:

$w_1 = -se^{-sR}t_1$, $w_2 = -se^{-sR}t_2$, $w_{11} = se^{-sR}Q_1$, $w_{22} = se^{-sR}Q_2$ and $w_{12} = se^{-sR}Q_3$ where $Q_1 = -(0.5)(s\theta_2\theta_1^{-1}R + 2)t_3$, $Q_2 = (0.5)(sR - 2)t_4$ and $Q_3 = \theta_2^{-1}t_1 + (0.5)sRt_3 - t_4$

The Bayes estimator for R, denoted by $\hat{R}_{BLL}$, under LINEX loss function can be evaluated by the form

$$\hat{R}_{BLL} = R - (1/s)ln(1 + sH). \tag{25}$$

where $H = H(\theta_1, \theta_2, \alpha) = (0.5)Q_1\sigma_{11} + (0.5)Q_2\sigma_{22} + Q_3\sigma_{12} - \Psi_1 t_1 - \Psi_2 t_2$,

Keeping in mind that these values are evaluated at the MLEs of θ_1, θ_2 and α.

5. Credible Interval

We know that the inference about R depends only on θ_1 and θ_2. However, the estimators of θ_1 and θ_2 depend on α, the estimation of R can be accomplshed as soon as α is estimated and become known. Depending on the ML estimate of α from the observed samples, we employed the empirical Bayesian procedure suggested by Lindely (1969) and used by Awad and Gharaf (1986). They had estimated the prior parameters of θ_1 and θ_2 empirically. From the likelihood function given in equation (5), one can see that $U = \sum_1^{n_1} ln(1 - e^{-x_i^\alpha})^{-1}$ and $V = \sum_1^{n_2} ln(1 - e^{-y_i^\alpha})^{-1}$ are sufficient statistics for θ_1 and θ_2, respectively. We have the assumption that θ_1 and θ_2 have independent gamma priors as $\theta_1 \sim G(a_1, b_1)$ and $\theta_2 \sim G(a_2, b_2)$. The empirical Bayes procedure suggests to take $a_1 = n_1 + 1$, $b_1 = U$, $a_2 = n_2 + 1$, $b_2 = V$ as estimated from the observed samples. When we adopt these empirical priors we get the posterior distributions $\theta_1|\underline{x} \sim G(a_1, b_1)$ and $\theta_2|\underline{y} \sim G(a_2, b_2)$ where $a_1 = 2n_1 + 1$, $b_1 = 2U$ and $a_2 = 2n_2 + 1$, $b_2 = 2V$. Therefore, we can get two independent chi-squared random variables Q_1 and Q_2 as $Q_1 = 4\theta_1 U \sim \chi^2(2N_1)$ and $Q_2 = 4\theta_2 V \sim \chi^2(2N_2)$, $N_1 = 2n_1 + 1$ and $N_2 = 2n_2 + 1$. The random variable $Q = (N_2 U\theta_1/N_1 V\theta_2) \sim F(2N_1, 2N_2)$, i.e. Q is F distributed random variable with $2N_1$ and $2N_2$ degrees of freedom. Hence, $Q = (N_2 U/N_1 V)(R/1 - R)$ can be used as a pivotal quantity to obtain a $100(1 - \tau)\%$ CrI for R. The lower and upper bounds of this interval can be obtained, respectively, as

$$L = F(2N_1, 2N_2; \tau/2)\Big[\frac{N_2 u}{N_1 v} + F(2N_1, 2N_2; \tau/2)\Big]^{-1}, U = F(2N_1, 2N_2; 1 - \tau/2)\Big[\frac{N_2 u}{N_1 v} + F(2N_1, 2N_2; 1 - \tau/2)\Big]^{-1}. \quad (26)$$

It is worth to mention that this interval performs very well in terms of its length compared with the confidence intervals in Section 3, as it is expected, when we apply to the real data as we will see in Section 6.

6. Data Analysis

For illustration purposes, we present a real data analysis of the strength of two types of data: (1) Single carbon fiber data and (2) Jute fiber data. We apply the estimation methods, presented here, for R.

(1) Single carbon fibers data

We present a real data analysis of the strength data reported by Badar and Priest (1982). The data represent the strength data measured in GPA (GigaPascal, $GPA = KN/mm2$, Kilonewten/squared millimeter, that it is used to measure tensile strength of materials such as nylon, fiber, . . .,etc.). We consider the data of single carbon fibers that were tested under tension at gauge lengths of 20 mm and 50 mm. The data sets are given as follows:

Data set 1 of length 20 mm: X ($n_1 = 69$)
1.312, 1.314, 1.479, 1.552, 1.700, 1.803, 1.861, 1.865, 1.944, 1.958, 1.966, 1.997, 2.006, 2.021, 2.027, 2.055, 2.063, 2.098, 2.140, 2.179, 2.224, 2.240, 2.253, 2.270, 2.272, 2.274, 2.301, 2.359, 2.382, 2.426, 2.435, 2.478, 2.490, 2.514, 2.535, 2.554, 2.566, 2.570, 2.586, 2.629, 2.633, 2.642, 2.648, 2.684, 2.697, 2.726, 2.773, 2.800, 2.809, 2.818, 2.821, 2.848, 2.880, 2.954, 3.012, 3.067, 3.084, 3.090, 3.096, 3.128, 3.233, 3.433, 3.585.
Data set 2 of length 50 mm: Y ($n_2 = 65$)
1.339, 1.434, 1.549, 1.574, 1.589, 1.613, 1.746, 1.753, 1.764, 1.807, 1.812, 1.840, 1.852, 1.852, 1.862, 1.864, 1.931, 1.952, 1.974, 2.019, 2.051, 2.055, 2.058, 2.088, 2.125, 2.162, 2.171, 2.172, 2.180, 2.194, 2.212, 2.270, 2.272, 2.280, 2.299, 2.308, 2.335, 2.349, 2.356, 2.386, 2.390, 2.410, 2.430, 2.431, 2.458, 2.471, 2.497, 2.514, 2.558, 2.577, 2.593, 2.601, 2.604, 2.620, 2.633, 2.670, 2.682, 2.699, 2.705, 2.735, 2.785, 3.020, 3.042, 3.116, 3.174.

Now we want to see whether the EW distribution can be used to fit these data sets or not. For this purpose we use the graphical approach called Q-Q plot for each data set. Q-Q plots are commonly used to compare a data set to a theoretical model. We construct the Q-Q plot by obtaining the points $(Q(i), x_i)$, $i = 1, 2, ..., m$ where $Q(i) = F^{-1}(i/(m + 1), \hat{\alpha}, \hat{\theta})$ and x_i is $i - th$ order statistic of the given data, $\hat{\alpha}$ and $\hat{\theta}$ are the MLE of α and θ. For the given data set 1 and 2, we get the MLEs $\hat{\alpha_1} = 1.4543$, $\hat{\theta_1} = 23.5641$ and $\hat{\alpha_2} = 1.6242$, $\hat{\theta_2} = 24.5255$, respectively. Hence, the shape parameters $\hat{\alpha_1}$ and $\hat{\alpha_2}$ of the distributions of the data sets are not very different. Therefore, the MLE, $\hat{\alpha}$, of common α is estimated to be 1.5224 and $\hat{\theta_1} = 27.0128$, $\hat{\theta_2} = 20.2886$. To support this claim, we also compute the log-likelihood values, $lnL_1(\underline{x}, \hat{\alpha_1}, \hat{\theta_1})$ and $lnL_2(\underline{y}, \hat{\alpha_2}, \hat{\theta_2})$ (in case of α is not common, $\alpha_1 \neq \alpha_2$), for the distribution of the two data sets, to find $lnL_1 = -52.3765$ and $lnL_2 = -36.4957$. In case of α is common ($\alpha_1 = \alpha_2 = \alpha$), we found that $lnL_1 = -53.0107$ and $lnL_2 = -37.4773$. These support that we cannot reject the null hypothesis that $\alpha_1 = \alpha_2$ and hence the claim that the two shape parameters for the distributions of thses data sets are equal, is justified. Figures 1 and 2 depict the Q-Q plots for both the data set 1 and 2. It is clear that the EW model fits quite well for both given data sets. This conclusion is also supported by the Kolmogrov-Smirnov (K-S) tests where the K-S statistic values are 0.0843 and 0.0929 with associated p values are 0.6784 and 0.5959, respectively.

Based on the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$, the ML estimate of R is $\hat{R}_M = 0.5711$ and the bootstrap estimate is $\hat{R}_{Boot} = 0.5721$. The ACI, p-boot CI and t-boot CI, with 95% confidence level, for R and their lengths are reported in Table 1. To evaluate

the Bayes estimates and credible interval, small values (0.001) for the hyper parameters of gamma prior densities were considered to the vague prior information allow to get meaningful comparison with MLE of R. From the Bayes estimators formulas in equations (23) and (25), the Bayes estimates of R is $\hat{R}_{BSL} = 0.5704$ and $\hat{R}_{BLL} = 0.5736$. We note that the estimated value of R is greater than 0.5, implying that the carbon fibers with length 20 mm is stronger than carbon fibers with length 50 mm. The 95% credible interval, CrI, for R, computed by the form given in the equation (26), and its length are reported in Table 1. Note that the CrI region is highly shorter in length than the corresponding confidence intervals. For bootstrap methods, the results are based on 5000 repeated samples.

Table 1. Confidence and credible intervals for R (single carbon fiber data)

ACI	p-boot CI	t-boot CI	CrI
(0.4880, 0.6541)	(0.4886, 0.6536)	(0.4885, 0.6536)	(0.4967,0.6142)
0.1661	0.1706	0.1663	0.1167

(2) Jute fibers data

These data sets are presented and studied by Xie et al. (2009). The data represent the breaking strength of Jute fiber at two different gauge lengths. The data sets are given as follows:

Data set 1 of length 10 mm: X ($n_1 = 30$):

693.73, 704.66, 323.83, 778.17, 123.06, 637.66, 383.43, 151.48, 108.94, 50.16, 671.49, 183.16, 257.44, 727.23, 291.27, 101.15, 376.42, 163.40, 141.38, 700.74, 262.90, 353.24, 422.11, 43.93, 590.48, 212.13, 303.90, 506.60, 530.55, 177.25.

Data set 2 of length 20 mm: Y ($n_1 = 30$):

71.46, 419.02, 284.64, 585.57, 456.60, 113.85, 187.85, 688.16, 662.66, 45.58, 578.62, 756.70, 594.29, 166.49, 99.72, 707.36, 765.14, 187.13, 145.96, 350.70, 547.44, 116.99, 375.81, 581.60, 119.86, 48.01, 200.16, 36.75, 244.53, 83.55.

To check whether the EW distribution can be used or not to fit these data sets, we use the Q-Q plot and K-S tests. The ML estimators for data sets 1 and 2 are $\hat{\alpha}_1 = 0.2703$ and $\hat{\alpha}_2 = 0.2703$, respectively, and hence the distributions of the two data sets have the same shape parameters $\alpha_1 = \alpha_2 = \alpha$. The ML estimate of the common shape parameter α is $\hat{\alpha} = 0.2681$ and hence $\hat{\theta}_1 = 62.1842$, $\hat{\theta}_2 = 48.8899$. The K-S statistic values are 0.1420 and 0.1376 with associated p values are 0.5341 and 0.5737, respectively. Therefore, one cannot reject the hypothesis that the data sets follow the EW distribution. Figures 3 and 4 show that the EW distribution fits well the tow data sets. For jute fiber data and under the same considerations for Bayes estimates (cited in case of single carbon fiber data), we get the following estimators of R: $\hat{R}_M = 0.5598$, $\hat{R}_{Boot} = 0.5693$, $\hat{R}_{BSL} = 0.5582$ and $\hat{R}_{BLL} = 0.5725$. We note that the estimated value of R is greater than 0.5, implying that the Jute fiber with length 10 mm is stronger than Jute fiber with length 20 mm. The ACI, p-boot CI and t-boot CI as well as CrI and their lengths are reported in Table 2. The results using the bootstrap methods are obtained over 5000 repeated samples.

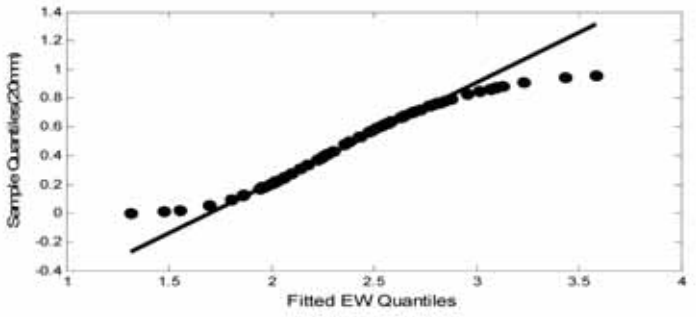


Figure 1. Q-Q plot of the fitted EW distribution for data set 1(single carbon fiber data)

Table 2. Confidence and credible intervals for R (jute fiber data)

ACI	p-boot CI	t-boot CI	CrI
(0.4448, 0.6944)	(0.4379, 0.6958)	(0.4434, 0.6958)	(0.4755,0.6491)
0.2496	0.2579	0.2524	0.1736

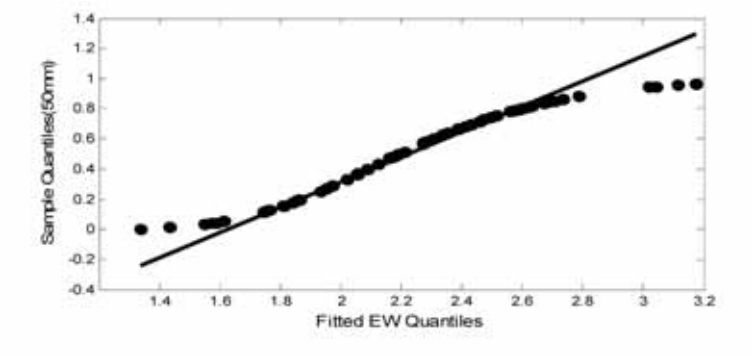


Figure 2. Q-Q plot of the fitted EW distribution for data set 2(single carbon fiber data)

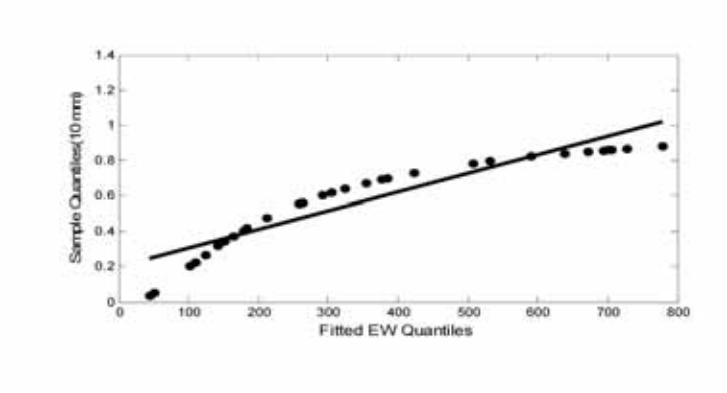


Figure 3. Q-Q plot of the fitted EW distribution for data set 1 (jute fiber data)

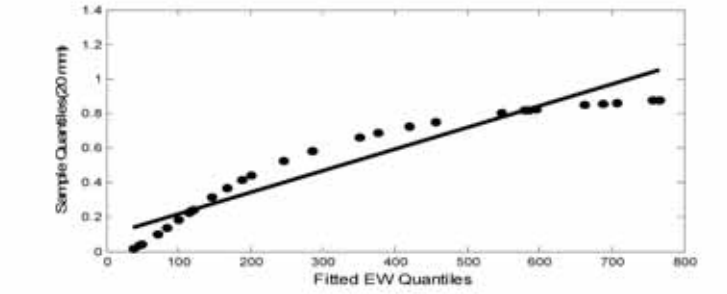


Figure 4. Q-Q plot of the fitted EW distribution for data set 2 (jute fiber data)

For the two data sets, the MLE and Bayes estimator (under non informative priors) perform quite similarly, while the length of the credible interval is the shortest compared with the corresponding confidence intervals obtained by other methods.

7. Simulation Study

A simulation study is carried out through some simulation experiments to see how the different estimation methods work for different values of $R = P(Y < X)$ using different sample sizes. We generate a set of 2000 X-samples from the $EW(\alpha, \theta_1)$ and another set of 2000 independent Y-samples from the $EW(\alpha, \theta_2)$. We choose the sample sizes n_1= 10, 20, 35 and 50 with combinations of the same values of n_2. The parameter values of α is 0.75 (1.5) with different several values of θ_1 and θ_2 to represent different values of the reliability parameter R to be 0.25, 0.40, 0.50, 0.70, 0.90. From the sample, we estimate α from equation (11) using a simple iterative algorithm. We employ the estimate of α to evaluate θ_1 and θ_2 using equations (10). Consequently, we get the MLE, $\hat{R}_M$ of R. For Bayesian estimation under squared error loss and LINEX loss functions, small values (0.001) for the hyper-parameters of gamma prior densities are considered to get meaningful comparison with MLE of R. We report the average mean squared errors (MSEs) of different estimators in Tables 3 and 4. We compute the 95% confidence interval based on asymptotic distribution of $\hat{R}_M$ and the bootstrap, p-boot and t-boot, confidence intervals as well as the credible interval. The average lengths and coverage probabilities (CPs) are reported for 95% confidence level in Tables 5 and 6.

From the results in Tables 3 and 4, some of points are observed from this simulation.

- All estimators perform quite well in terms of the MSEs for all sample sizes.

- The ML estimator works well even with small sample size. This show that the coincidence and consistency properties of all estimators.

- The MSE of $\hat{R}_{BSL}$ is the smallest comparing with that of the other estimators, especially for small sample sizes.

- The MSEs decrease as the sample size increases for all methods and for different values of R.

- For the same size of the samples (say, for samples in sizes (10,10) or in sizes (35,20) at different values of R), the MSEs increase when $0 < R \leq 0.5$ and decrease when $0.5 < R \leq 1$ as R value increases through these two ranges.

- For small sample sizes, the MSEs of the different estimators in case of $n_1 \neq n_2$ is smaller than the MSEs in case of $n_1 = n_2$.

Examining Tables 5 and 6, it is clear that:

- The average lengths of all intervals decrease as the sample size increases.

- The average lengths of the credible interval are smaller than that of the asymptotic and bootstrap confidence intervals for all different values of R and different sample sizes.

- For the same size of the samples, at the values of R, $0 < R \leq 0.5$, the increasing values of R the increasing the average lengths of different intervals and conversely when $0.5 < R \leq 1$.

- For small sample sizes, the average lengths of the different intervals in case of $n_1 \neq n_2$ is smaller than the lengths in case of $n_1 = n_2$.

- The coverage probabilities of the bootstrap confidence intervals are able to preserve the nominal level even for small sample sizes.

- The coverage probabilities of the asymptotic confidence intervals are slightly lower than the nominal level.

- The coverage probabilities of the credible intervals based on lack information a priori, are lower than the nominal level.

- In brief, the performances of the bootstrap confidence intervals are the best among the intervals taken into account here. Also, the credible interval is the best in terms of the lengths of the intervals.

Other simulation results were also considered at $\alpha = 1.5$ for the same sample sizes cited above. The results are not reported here since they have a similar pattern to the results in Tables 3, 4, 5 and 6.

8. Conclusion

In this article, we studied the Bayesian and non Bayesian Inferences of the stress-strength parameter $R = P(X > Y)$ when X and Y both follow the exponentiated Weibull distribution. We employed the ML method to estimate the MLE of R. The exact distribution of R is difficult to obtain and then we resorted to use the asymptotic distribution to compute the asymptotic confidence interval. Parametric bootstrap procedure is conducted and evaluate the estimate of R as well as different bootstrap confidence intervals are computed. We derived two Bayes estimates of R based on the independent gamma priors, using the approximate Lindely's procedure under squared error loss and LINX loss functions. Also, we derived the credible interval using the empirical method of Lindely (1969) and Awad and Gharaf (1986). The simulation results indicate that the Bayesian estimator under squared error loss function works the best even for small sample sizes.

The credible intervals perform the best in terms of the average lengths of the intervals in both cases of $n_1 \neq n_2$ and $n_1 = n_2$ of the samples. The bootstrap confidence intervals are the best in terms of the nominal level taken into account in the simulation. Using real data, we examine the different estimations over two actual data sets.

Table 3. MSEs for different estimates of R, $n_1 = n_2$

R	(n_1, n_2)	$\hat{R}_M$	$\hat{R}_{Boot}$	$\hat{R}_{BSL}$	$\hat{R}_{BLL}$
0.25	(10,10)	0.0072	0.0074	0.0070	0.0069
	(20,20)	0.0036	0.0038	0.0036	0.0034
	(35,35)	0.0019	0.0020	0.0019	0.0019
0.40	(10,10)	0.0108	0.0120	0.0095	0.0107
	(20,20)	0.0064	0.0059	0.0060	0.0063
	(35,35)	0.0030	0.0033	0.0029	0.0029
0.50	(10,10)	0.0125	0.0130	0.0109	0.0129
	(20,20)	0.0061	0.0062	0.0057	0.0062
	(35,35)	0.0034	0.0036	0.0033	0.0035
0.70	(10,10)	0.0095	0.0096	0.0089	0.0097
	(20,20)	0.0045	0.0046	0.0043	0.0045
	(35,35)	0.0025	0.0025	0.0025	0.0025
0.90	(10,10)	0.0020	0.0020	0.0021	0.0020
	(20,20)	0.0008	0.0010	0.0009	0.0008
	(35,35)	0.0005	0.0006	0.0005	0.0005

Table 4. MSEs for different estimates of R, $n_1 \neq n_2$

R	(n_1, n_2)	$\hat{R}_M$	$\hat{R}_{Boot}$	$\hat{R}_{BSL}$	$\hat{R}_{BLL}$
0.25	(10,20)	0.0072	0.0066	0.0070	0.0059
	(35,20)	0.0023	0.0027	0.0030	0.0028
	(35,50)	0.0018	0.0018	0.0018	0.0018
0.40	(10,20)	0.0092	0.0093	0.0083	0.0091
	(35,20)	0.0047	0.0042	0.0045	0.0046
	(35,50)	0.0028	0.0029	0.0027	0.0028
0.50	(10,20)	0.0081	0.0096	0.0073	0.0083
	(35,20)	0.0049	0.0052	0.0046	0.0050
	(35,50)	0.0030	0.0028	0.0029	0.0030
0.70	(10,20)	0.0067	0.0071	0.0043	0.0068
	(35,20)	0.0038	0.0034	0.0036	0.0038
	(35,50)	0.0018	0.0023	0.0018	0.0019
0.90	(10,20)	0.0015	0.0014	0.0018	0.0015
	(35,20)	0.0008	0.0008	0.0008	0.0008
	(35,50)	0.0005	0.0005	0.0005	0.0005

Table 5. Average lengths for different intervals (CPs in brackets), $n_1 = n_2$

R	(n_1, n_2)	ACI	CP	p-boot	CP	t-boot	CP	CrI	CP
0.25	(10,10)	0.3221	(0.9045)	0.3363	(0.9571)	0.3599	(0.9695)	0.2526	(0.8780)
	(20,20)	0.2317	(0.9235)	0.2413	(0.9505)	0.2596	(0.9680)	0.1813	(0.9120)
	(35,35)	0.1770	(0.9420)	0.1732	(0.9492)	0.1771	(0.9505)	0.1367	(0.9000)
0.40	(10,10)	0.4041	(0.9180)	0.4234	(0.9485)	0.4364	(0.9555)	0.3086	(0.9120)
	(20,20)	0.2916	(0.9280)	0.2974	(0.9507)	0.3000	(0.9520)	0.2319	(0.9040)
	(35,35)	0.2226	(0.9390)	0.2212	(0.9506)	0.2249	(0.9560)	0.1733	(0.9160)
0.50	(10,10)	0.4163	(0.9070)	0.4511	(0.9535)	0.4437	(0.9495)	0.3231	(0.9060)
	(20,20)	0.3025	(0.9340)	0.3044	(0.9512)	0.2966	(0.9400)	0.2344	(0.9161)
	(35,35)	0.2311	(0.9380)	0.2319	(0.9532)	0.2323	(0.9505)	0.1790	(0.9240)
0.70	(10,10)	0.3588	(0.9051)	0.3761	(0.9537)	0.3632	(0.9395)	0.2816	(0.9000)
	(20,20)	0.2600	(0.9230)	0.2635	(0.9515)	0.2542	(0.9405)	0.2027	(0.8940)
	(35,35)	0.1981	(0.9235)	0.1910	(0.9505)	0.1911	(0.9505)	0.1532	(0.9180)
0.90	(10,10)	0.1663	(0.9000)	0.1605	(0.9405)	0.1318	(0.9100)	0.1325	(0.8660)
	(20,20)	0.1195	(0.9220)	0.1213	(0.9507)	0.1084	(0.9255)	0.0903	(0.8911)
	(35,35)	0.0903	(0.9330)	0.0934	(0.9490)	0.0857	(0.9365)	0.0676	(0.9021)

Table 6. Average lengths for different intervals (CPs in brackets), $n_1 \neq n_2$

R	(n_1, n_2)	ACI	CP	p-boot	CP	t-boot	CP	CrI	CP
0.25	(10,20)	0.2877	(0.9265)	0.3082	(0.9505)	0.3671	(0.9745)	0.2288	(0.8800)
	(35,20)	0.2062	(0.9245)	0.1980	(0.9481)	0.1984	(0.9515)	0.1628	(0.8900)
	(35,50)	0.1632	(0.9355)	0.1615	(0.9511)	0.1790	(0.9690)	0.1312	(0.9000)
0.40	(10,20)	0.3539	(0.9190)	0.3658	(0.9478)	0.4004	(0.9680)	0.2800	(0.8740)
	(35,20)	0.2595	(0.9320)	0.2566	(0.9496)	0.2459	(0.9375)	0.2055	(0.8820)
	(35,50)	0.2056	(0.9360)	0.2129	(0.9506)	0.2225	(0.9600)	0.1623	(0.9040)
0.50	(10,20)	0.3652	(0.9120)	0.3721	(0.9477)	0.4103	(0.9700)	0.2813	(0.8920)
	(35,20)	0.2697	(0.9375)	0.2788	(0.9502)	0.2680	(0.9370)	0.2117	(0.8920)
	(35,50)	0.2134	(0.9370)	0.2078	(0.9501)	0.2119	(0.9545)	0.1681	(0.9100)
0.70	(10,20)	0.3124	(0.8970)	0.3169	(0.9488)	0.3379	(0.9620)	0.2421	(0.8480)
	(35,20)	0.2314	(0.9371)	0.2273	(0.9510)	0.2084	(0.9340)	0.1815	(0.9100)
	(35,50)	0.1832	(0.9375)	0.1841	(0.9503)	0.1899	(0.9560)	0.1380	(0.9200)
0.90	(10,20)	0.1424	(0.8835)	0.1372	(0.9455)	0.1350	(0.9480)	0.1110	(0.8220)
	(35,20)	0.1050	(0.9385)	0.1072	(0.9485)	0.0879	(0.9000)	0.0832	(0.9000)
	(35,50)	0.1462	(0.9230)	0.1534	(0.9490)	0.1148	(0.8840)	0.0647	(0.8540)

References

Awad, A. M., Azzam, M. M., & Hamadan, M. A. (1981). Some inferences results in $P(Y < X)$ in the bivariate exponential model. *Communications in Statistics- Theory and Methods, 10*, 2515-2525.

Badar, M. G., & Priest, A. M. (1982). Statistical aspects of fiber and bundle strength in hybrid composites. *Progress in Science and Engineering Composites.* In: Hayashi, T., Kawata, K., Umekawa, S., eds. Progress in Science and Engineering Composites. Tokyo: ICCM-IV, 1129-1136.

Birnbaum, Z. W. (1956). On a use of Mann-Whitney Statistics. Proceeding Third Berkeley Symposium on *Mathematical Statistics and Probability, 1*, 13-17.

Constantine, K., Tse, S-K., & Karson, M. (1986). Estimation of $P(Y < X)$ in gamma case. *Communications in Statistics - Computations and Simulations, 15*, 365- 388. https://doi.org/10.1080/03610918608812513.

Efron, B., & Tibshirani, R. (1998). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 0-412-14321-2.

Hauk, W. W., Hyslop, T., & Anderson, S. (2000). Generalized treatment efects for clinical trials. *Statistics in Medicine, 19*, 887-899. https://doi.org/10.1002/(SICI)1097-0258.

Jovanovic, M., & Rajic, V. (2014). Estimation of $P(X < Y)$ for Gamma Exponential Model. *Yugoslav Journal of*

operations Research, 2, 283-291. https://doi.org/10.2298/YJOR121020006J.

Lindely, D. V. (1969). *Introduction to Probability and Statistics from a Bayesian Viewpoint*, *1*. Cambridge University Press, Cambridge.

Lindely, D. V. (1980). Approximate Bayes Method. *Trabajos de Estadistica, 3*, 281-288.

Mahdizadeh, M. (2018). On estimating a stress-strength type reliability. *Hacettepe Journal of Mathematics and Statistics, 47*(1), 243-253. https://doi.org/10.15672/HJMS.201612418374.

Mudholkar, G. S., & Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability, 42*(2), 299-302. https://doi.org/10.1109/24.229504.

Raqab, M. Z., Madi, M. T., & Kundu, D. (2008). Estimation of $P(Y < X)$ for the three-parameter generalized exponential distribution. *Communications in Statistics- Theory and Methods, 37*, 2854-2864. https://doi.org/10.1080/03610920802162664.

Rao, C. R. (1973). *Linear Statistical Inference and its Application*. Second Edition. John Willy & Sons, New York. https://doi.org/10.1002/zamm.19770570832.

Rao, G. S., Rosaiah, K., & Babu, M. S. (2016). Estimation of Stress-Strength reliability from exponentiated Frechet Distribution. *International Journal of Advanced Manufacturing Technology, 86*, 3041-3049. https://doi.org/10.1007/s00170-016-8404-z.

Reiser, B. (2000). Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. *Statistics in Medicine, 19*, 2115-2129. https://doi.org/10.1002/1097-0258.

Rezaei, S., Tahmasbi, R., & Mahmoodi, M. (2010). Estimation of $P(Y < X)$ for the generalized Pareto distribution. *Journal of Statistical Planning and Inference, 140*, 480-494. https://doi.org/10.1016/j.jspi.2009.07.024.

Sarhan, A. M., Smith, B., & Hamilton, D. C. (2015). Estimation of $P(Y < X)$ for a Two-parameter Bathtub Shaped Failure Rate Distribution. *International Journal of Statistics and Probability, 4*(2), 33-45. https://doi.org/10.5539/ijsp.v4n2p33.

Weerahandi, S., & Johnson, R. A. (1992). Testing reliability in stress-strength model when X and Y are normally distributed. *Technometrics, 34*, 83- 91. https://doi.org/10.1080/00401706.1992.10485236.

Wellek, S. (1993). Basing the analysis of comparative bioavailability trials on an individualized statistical definition of equivalence. *Biometrical Journal, 35*, 47-55. https://doi.org/10.1002/bimj.4710350105.

Xie, Z. P., Yu, J. Y., Cheng, L. D., Liu, L. F., & Wang, W. M. (2009). Study on the breaking strength of jute fibers using modified Weibull distribution. *Journal of Composites Part A: Applied Science and Manufactoring, 40*, 54-59. https://doi.org/10.1016/j.compositesa.2008.10.001.

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The Bayes Factor for the Misclassified Categorical Data

Tze-San Lee[1]

[1] Western Illinois University, USA

Correspondence: Tze-San Lee, retired mathematical professor at Western Illinois University.

Received: May 2, 2018 Accepted: May 23, 2018 Online Published: June 28, 2018

doi:10.5539/ijsp.v7n4p91 URL: https://doi.org/10.5539/ijsp.v7n4p91

Abstract

This article addresses the issue of misclassification in a single categorical variable, that is, how to test whether the collected categorical data are misclassified. To tackle this issue, a pair of null and alternative hypotheses is proposed. A mixed Bayesian approach is taken to test these hypotheses. Specifically, a bias-adjusted cell proportion estimator is presented that accounts for the bias caused by classification errors in the observed categorical data. The chi-square test is then adjusted accordingly. To test the null hypothesis that the data are not misclassified under a specified multinomial distribution against the alternative hypothesis they are misclassified, the Bayes factor is calculated for the observed data and a comparison is made with the classical p-value.

Keywords: Bayes factor, classification errors, Dirichlet's distribution, Type II maximum likelihood

1. Introduction

The problem of misclassification is a major issue in observational epidemiologic studies. Not long after Bross (1954) pointed out that the non-differential misclassification would bias the corrected odds ratio toward the null hypothesis, Diamond and Lilienfeld (1962a-b) has extended the result to various types of epidemiologic studies. A 2 × 2 case-control studies with a single exposure variable being misclassified has been widely studied (Fleiss et al 2003, Chapter 17; Gustafson 2004, Chapter 5; Kleinbaum et al 1982, Chapter 12; Rothman et al 2008, Chapter 19). Yet, almost no authors pay attention to investigate the effect of misclassification in the analysis of a single categorical variable except Mote and Anderson (1965). Mote and Anderson primarily takes a deductive approach to account for the bias caused by the classification errors. Yet, the shortcoming with a deductive approach is that it does not take the sampling errors into consideration. As a result, the issue on how to deal with the misclassification in the analysis of categorical data still remains unsolved.

This article addresses another important issue, that is, whether the observed categorical data are misclassified. Instead of using a deductive method, an inductive approach is employed to account for the misclassification bias embedded in the collected data. First, the inverse way is taken by equating the expected value of the estimated sample cell proportion with its population parameter conditional on that the misclassification probabilities are given. Then the bias-adjusted estimator is presented for the population cell proportion parameter by inverting the misclassification matrix. Second, the appropriate misclassification probabilities are calculated depending on if the misclassification is possibly made either from one category to all other categories (scenario I) or merely to its neighboring categories (scenario II). Third, in order to test the null hypothesis that the data are not misclassified under a specified multinomial distribution, a mixed Bayesian approach is used to calculate the Bayes factor and compare it with the traditional p-value.

2. Methodology & Background

Given that X is a categorical variable with K (≥ 3) categories and the data are collected through a simple random sampling of size N, where $N = \sum_{i=1}^{K} n_i$ (table 1). The crude estimator, $\hat{p}_j$, for the population cell proportion p_j in the j^{th} category is then given by

$$\hat{p}_j = n_j / N . \tag{1}$$

Assume that $\hat{p}_j$ is distributed as a multinomial distribution with the population size N and the cell proportion of the j^{th} category p_j. It is well known that Eq. 1 is an unbiased estimator for the population cell proportion parameter, provided that the observed data are not misclassified (Agresti 2002). However, it is shown below by Eq. 4 that $\hat{p}_j$ of Eq. 1 is no longer unbiased for p_j, once the observed data are misclassified.

Table 1. Observed data for the categorical variable X

Variable	Categories			
X	1	2	……	K
Observation	n_1	n_2	……	n_K

Suppose that the observed data are misclassified. Let w_{jk} ($j \neq k$) be the misclassification probability of an observation belonging to the j^{th} category being incorrectly classified into the k^{th} category and w_{jj} the correct classification probability that an observation belonging to the j^{th} category being correctly classified into the j^{th} category. Then, it is easily shown that the expected value of $\hat{p}$ is

$$E(\hat{p}) = Wp, \tag{2}$$

where $p = (p_1, p_2, ..., p_K)$, $\hat{p} = (\hat{p}_1, \hat{p}_2, ..., \hat{p}_K)$, and $W = [w_{jk}]^T{}_{j,k=1,2,...,K}$ is the misclassification matrix, in which $\sum_{k=1}^{K} w_{jk} = 1$ for j = 1, 2, …, K. Eq. 2 shows that the crude estimator $\hat{p}_k$ is no longer unbiased for the population parameter p_k, provided that $W \neq I$, where I is the K × K identity matrix. A set of misclassification probabilities $\{w_{jk}\}$ is said to be feasible if the misclassification matrix W in Eq. 2 is invertible (or nonsingular) for $0 < w_{jk} < 1$.

Assume that W is invertible. Then bias-adjusted cell proportion (BACP) estimators ($\breve{p}_k$) are defined by

$$\breve{p} = W^{-1}\hat{p} = V\hat{p}, \tag{3}$$

where $\breve{p} = (\breve{p}_1, \breve{p}_2, ..., \breve{p}_K)^T$, V = [$v_{jk}$], j, k = 1, 2, …, K, denotes the inverse matrix of W, and $\breve{n} = Vn$, $\breve{n} = (\breve{n}_1, ,,,, \breve{n}_K)^T$, $n = (n_1, ..., n_K)^T$. Note that by using Eqs. 2 and 3 it's easily shown: $E(\breve{p}) = p$, namely, $\breve{p}$ is an unbiased estimator for p, provided that W is known. The BACP estimators $\{\breve{p}_k\}$ are said to be admissible if for feasible w_{jk} we have $0 < \breve{p}_k < 1$ and $\sum_{j=1}^{K} \breve{p}_j = 1$. Similarly, a set of misclassification error probabilities $\{w_{jk}\}$ is said to be admissible if the corresponding BACP estimators $\{\breve{p}_k\}$ are admissible.

The misclassification matrix W has two possible forms depending on how the categorical variable X is misclassified. There are two possible scenarios that are given as follows:

Scenario I: The misclassification occurs after classifying one category incorrectly into all other categories. Also, because misclassification can occur equally likely from any one of the j^{th} correct category to the k^{th} (observed) wrong category, we thus have, for fixed j

$$\theta_j \equiv w_{jk} > 0, \; k \neq j, \text{ and } \; w_{jj} = 1 - \sum_{\substack{k=1 \\ k \neq j}}^{K} w_{jk}, \; j = 1, 2, \ldots, K, \tag{4}$$

Scenario II: The misclassification occurs after classifying one category incorrectly only into its neighboring categories. Therefore, we have, for fixed j

$$w_{jk} = 0 \text{ for } |k - j| > 1, \text{ and } \; w_{jj} = 1 - \sum_{\substack{k=1 \\ k \neq j}}^{K} w_{jk}, \; j = 1, 2, \ldots, K. \tag{5}$$

When K = 3, the associated misclassification matrix with its determinant and its inverse matrix for scenarios I and II are hereby obtained respectively. An explicit form of the misclassification matrix W_I and its inverse V_I for scenario I are given respectively by

$$W_I = \begin{bmatrix} 1-\theta_2-\theta_3 & \theta_2 & \theta_3 \\ \theta_1 & 1-\theta_1-\theta_3 & \theta_3 \\ \theta_1 & \theta_2 & 1-\theta_1-\theta_2 \end{bmatrix}, \tag{6a}$$

$$\Delta_I \equiv \det(W_I) = (1-\theta_1-\theta_2-\theta_3)^2 \neq 0, \tag{6b}$$

and

$$V_I \equiv [v_{jk(I)}] = \Delta_I^{-\frac{1}{2}} \cdot \begin{bmatrix} 1-\theta_1 & -\theta_1 & -\theta_1 \\ -\theta_2 & 1-\theta_2 & -\theta_2 \\ -\theta_3 & -\theta_3 & 1-\theta_3 \end{bmatrix}, \tag{6c}$$

where $\theta_1 \equiv w_{12} = w_{13}$, $\theta_2 \equiv w_{21} = w_{23}$, and $\theta_3 \equiv w_{31} = w_{32}$.

The BACP estimators for scenario I are given by

$$\breve{p}_{k(I)} = \sum_{j=1}^{K} v_{jk(I)} \cdot \hat{p}_j \ , \quad \text{k = 1, 2, …, K,} \tag{7}$$

By using Eqs. 6b and 7, the feasibility and admissibility constraints for the misclassification probability and BACP estimator are given respectively as follows:

$$\theta_1 + \theta_2 + \theta_3 < 1, \tag{8a}$$

and

$$\theta_1 < 1, \qquad \theta_2 < 1, \qquad \theta_3 < 1. \tag{8b}$$

For scenario II, an explicit form of the misclassification matrix W_{II} and its inverse V_{II} are given respectively by

$$W_{II} = \begin{bmatrix} 1-\gamma_2 & \gamma_2 & 0 \\ \gamma_1 & 1-\gamma_1-\gamma_3 & \gamma_3 \\ 0 & \gamma_2 & 1-\gamma_2 \end{bmatrix}, \tag{9a}$$

$$\Delta_{II} \equiv \det(W_{II}) = (1-\gamma_2)(1-\gamma_1-\gamma_2-\gamma_3) \neq 0, \tag{9b}$$

and

$$V_{II} \equiv [v_{jk(II)}] = \Delta_{II}^{-1} \cdot \begin{bmatrix} (1-\gamma_1)(1-\gamma_2) & -\gamma_1(1-\gamma_2) & \gamma_1\gamma_2 \\ -\gamma_2(1-\gamma_2) & (1-\gamma_2)^2 & -\gamma_2(1-\gamma_2) \\ \gamma_2\gamma_3 & -\gamma_3(1-\gamma_2) & (1-\gamma_2)(1-\gamma_3)-\gamma_1 \end{bmatrix}, \tag{9c}$$

where $\gamma_1 \equiv w_{12}$, $\gamma_2 \equiv w_{21} = w_{23}$, $\gamma_3 \equiv w_{32}$, and $w_{13} = w_{31} \equiv 0$.

The BACP estimator for scenario II is thus given by

$$\breve{p}_{j(II)} = \sum_{k=1}^{K} v_{jk(II)} \cdot \hat{p}_k \ , \quad \text{j = 1, 2, …, K.} \tag{10}$$

By using Eqs. 9b and 10, the feasibility and admissibility constraints for the misclassification probability and BACP are given respectively as follows:

$$\gamma_1 + \gamma_2 + \gamma_3 < 1, \tag{11a}$$

and

$$\gamma_2 < \hat{p}_2. \tag{11b}$$

To test whether the data in table 1 are misclassified, we need to test the following (sharp) null hypothesis that the data has no misclassification under $p = p^0$ versus the alternative hypothesis that the data are misclassified (Berger and Selleke

1987)

$$H_0: p = p^0, \omega = 0 \text{ versus } H_1: p \neq p^0, \omega > 0, \tag{12}$$

where $p = (p_1,...,p_K)^T$, $p^0 = (p_1^0,...,p_K^0)^T$, $\omega = (w_{11},...,w_{1K},w_{21},...,w_{2K},...,w_{K1},...w_{KK})^T$, $\{w_{jk}\}$ are the entries of the misclassification matrix W given by Eq. 2.

To test Eq. 12 the bias-adjusted chi-square test (BACST) is given by

$$\breve{\Psi}_K = \sum_{k=1}^{K} N[(\breve{p}_k - p_k^0)^2 / p_k^0] = \sum_{k=1}^{K} (\breve{n}_k^2 / n_k^0) - N , \tag{13}$$

where $\breve{n}_k = \sum_{j=1}^{K} v_{jk} n_j$, v_{jk} denotes the entry of the j^{th} row and the k^{th} column of the inverse matrix V of the misclassification matrix W in Eq. 2 and $n_k^0 = Np_k^0$, k = 1,…, K.

For large samples, Eq. 13 is distributed under H_0 asymptotically as the central chi-square distribution with K – 1 degrees of freedom (df). Yet Eq. 13 is distributed asymptotically under H_1 as the noncentral chi-square distribution with K – 1 degrees of freedom and the non-centrality parameter given by (Lancaster 1969)

$$\breve{\lambda}_K = \sum_{j=1}^{K} (p_j - p_j^0)^2 = \sum_{j=1}^{K} (p_j^2 - 2p_j^0 p_j + p_j^{02}) . \tag{14}$$

When $w_{jk} = 0$ for all j and k, Eq. 13 reduces to

$$\hat{\Psi}_K = \sum_{j=1}^{K} (n_j^2 / n_j^0) - N . \tag{15}$$

Reject the null hypothesis H_0 if $\hat{\Psi}_K \geq C_0$, where $\hat{\Psi}_K$ is given by Eq. 15 and C_0 is the critical value of the central chi-square distribution with K – 1 df at the significance level α

As is well known from the Bayesian viewpoint, the p-value is not an adequate measure for the evidence to support the null hypothesis (Goodman 1999a-b). Hence the Bayes factor is calculated as a comparison with the p-value. To formulate the hypothesis-testing problem in a Bayesian setting we begin with the data $n = (n_1, n_2,..., n_K)$ and assume that its probability distribution follows in a family of distributions which are parameterized by $(p, \omega) \in \Sigma \times \Omega$, where $\Sigma = \{p \mid \sum_{k=1}^{K} p_k = 1, p_k > 0\}$ is the K-dimensional simplex. To test the hypotheses of $H_0 : p = p^0, \omega = 0$ vs $H_1 : p \neq p^0, \omega > 0$ (Eq.12), it is assumed that there exist a prior probability density function (PDF) $h_0(\omega)$ and another joint density $h(p, \omega)$ under H_1. Since p and ω are a priori independent under H_1, we have

$$h(p, \omega) = h_0(\omega) g(p) , \tag{16}$$

where g is a prior PDF on p ϵ Σ which assigns mass π_0 to $\{p = p^0\}$ and $1 - \pi_0$ to $\{p \neq p^0\}$. Define $g(p^0) = 0$ and writing the PDF of $\breve{\Psi}_K$ given p and ω as $f(\breve{\Psi}_K \mid p, \omega)$, the Bayes factor is given by (Kass and Raftery 1995)

$$B^g(\breve{\Psi}_K) = \frac{f(\breve{\Psi}_K \mid p^0, \omega = 0)}{m_g(\breve{\Psi}_K)} , \tag{17a}$$

where m_g is given by

$$m_g(\breve{\Psi}_K) = \iint_{\Sigma\times\Omega} f(\breve{\Psi}_K \mid p,\omega) h_0(\omega) g(p) d\omega dp \,. \tag{17b}$$

In Eq. 17a, $f(\breve{\Psi}_K \mid p^0, \omega = 0)$ is the PDF of the central chi-square distribution with K – 1 df, while $f(\breve{\Psi}_K \mid p,\omega)$ in Eq. 17b is the PDF of the noncentral chi-square distribution with K – 1 degrees of freedom and the non-centrality parameter $\breve{\lambda}_K$ given by Eq. 14.

When K = 3, $m_g(\breve{\Psi}_{3(I)})$ of Eq. 17b is calculated for Scenario I with the assumption of $\theta_1 = \theta_2 = \theta_3 \equiv \theta$ and $h_0(\theta) = c^{-1}$, the PDF of uniform distribution over [0, c], where c is the upper bound on the admissible BACP for scenario I and obtain

$$m_g(\breve{\Psi}_{3(I)}) = \int_{\Sigma}\int_{0}^{c} \frac{1}{c} \cdot \frac{1}{2+\breve{\lambda}_3} \cdot \exp(-\frac{t}{2+\breve{\lambda}_3}) d\theta \cdot g(p) dp \,, \tag{18}$$

where an approximation to the noncentral chi-square distribution is provided by using the central chi-square distribution (Cox and Reid 1987). The lower bound for the Bayes factor after using a symmetric Dirichlet's prior for g(p) are obtained under scenario I and II:

$$\underline{B}_i^g = \frac{\frac{1}{2}\exp\{-\frac{1}{2}[\sum_{j=1}^{3}(n_j^2/n_j^0) - N]\}}{m_g(\tau_{\max(i)} \mid \breve{\Psi}_{3(i)})} \quad \text{, i = I or II.} \tag{19}$$

The details for obtaining the value of $\tau_{\max(i)}$, i = I or II, are given in the appendix.

3. Example

The data in table 2 are taken from table C.1 in Woodward's book, pp. 756-760 (Woodward 2005). It represents the lung cancer data collected by the Bombay Cancer Registry from all cancer patients registered in the 168 government and private hospitals and nursing homes in Bombay, Australia, and from death records maintained by the Bombay Municipal Corporation. The survival times of each subject with lung cancer from time of first diagnosis to death (or censoring) were recorded over the period 1st January 1989 to 31st December 1991. Here we are only concerned with type of tumor of 682 subjects grouped by gender.

Table 2. 682 cancer patients are classified by sex and type-of-tumor

	Type of tumor			
Gender	Local	Regional	Advanced	Total
Male	165	169	229	563
Female	37	39	43	119

The issue of concern here is whether the data are misclassified separately for males and females. Because we do not have any prior belief on the values of p^0 in Eq. 12, they are thereby determined empirically from the observed data. As a result, the values of p^0 are chosen differently for males and females. For females the values of p^0 in the null hypothesis are chosen to be that of equiprobability, $H_{0(F)} : p_1 = p_2 = p_3 = \frac{1}{3}$ and $w_{jk} = 0$ vs $H_{1(F)} : p_1 \neq p_2 \neq p_3 \neq \frac{1}{3}$ and $w_{jk} > 0$, while that of p^0 in the null hypothesis for males are set up as follows: $H_{0(M)} : p_1 = 0.3, p_2 = 0.3, p_3 = 0.4$ and $w_{jk} = 0$

vs $H_{1(M)} : p_1 \neq 0.3, p_2 \neq 0.3, p_3 \neq 0.4$ and $w_{jk} > 0$. Because the misclassification probabilities of $\{w_{jk}\}$, j, k = 1, 2, 3 are zero under the null hypothesis, the BACST values of Eq. 15 are then given respectively by $\hat{\Psi}_M$ = 0.15 (p-value = 0.93) and $\hat{\Psi}_F$ = 0.47 (p-value = 0.79) for males and females. Therefore, the null hypothesis H_0 is not rejected at the significance level of 0.05 for both males and females. Yet, we would like to test the above hypotheses from the Bayesian perspective by calculating the Bayes factor as a comparison with the p-value.

For both males and females under scenarios I or II, Eq. A10 in the appendix has three negative and one positive real, and a pair of conjugate complex roots. Due to the constraint that $\tau > 0$, only the positive root is a stationary point for Eq. A9. Eq. A9 for males has only under scenario II a unique positive local maximum (Figure 1), while Eq. A9 has a unique positive local maximum at its stationary point for females only under scenario I (Figure 2).

Table 3. A comparison of the lower bound for Bayes factor (Eq. 19) with the p-value for admissible CF models

Scenario II					
Males	c_2	$\tau_{\max(II)}$	$m_g(\tau_{\max(II)})$	$\underline{B}^g_{II}$	p-value
Table 2	0.293073	0.0553	61	0.053	0.93
Scenario I					
Females	c_1	$\tau_{\max(I)}$	$m_g(\tau_{\max(I)})$	$\underline{B}^g_{I}$	p-value
Table 2	0.310924	0.0540	1.8	0.22	0.79

By taking the reciprocal of the lower bound of the Bayes factor (table 3, column 5) we are able to assess the evidence whether the cancer data in table 2 are misclassified. The collected data for males were in favor of supporting H_1 against H_0 by at most a factor of "19 to 1", whereas for females by at most a factor of "5 to 1".

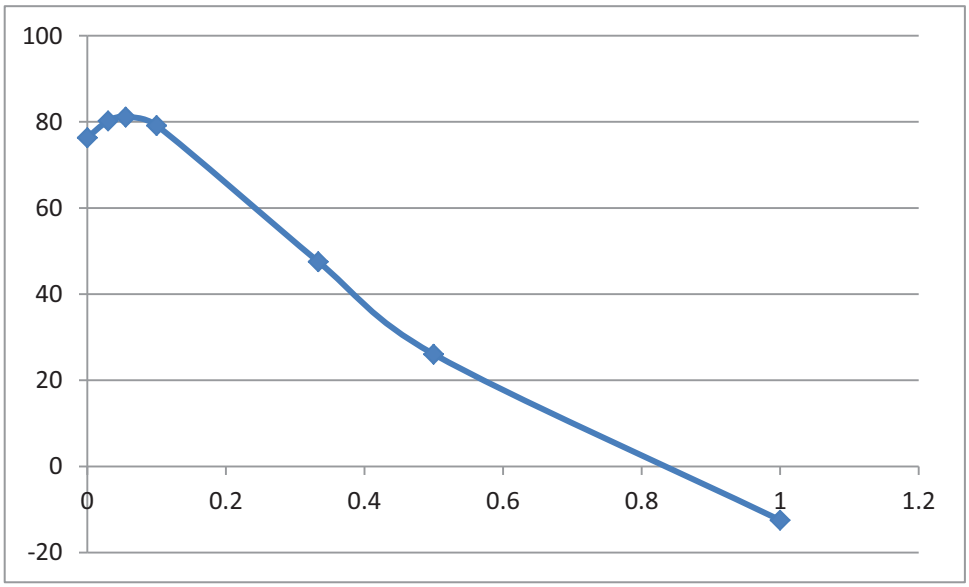


Figure 1. A plot of $m_g(\tau \mid \dot{\Psi}_{3(II)})$ given by Eq. A9 is for CF model 10 under scenario II for males

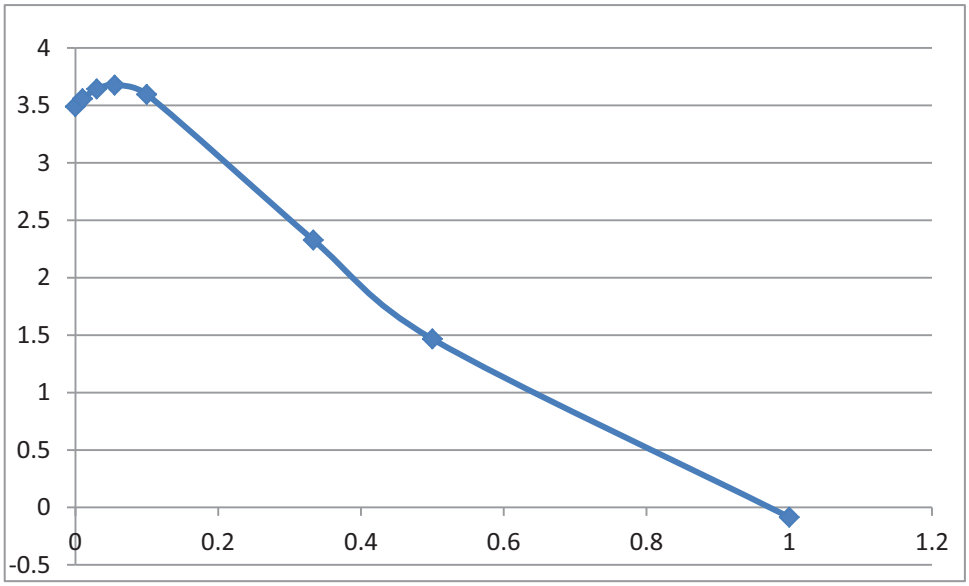


Figure 2. A plot of $m_g(\tau \mid \dot{\Psi}_{3(I)})$ given by Eq. A9 is for CF model 12 under scenario I for females

4. Discussion

Some interesting observations are worthy to be mentioned below:

1. So far, this author is not aware of any guideline available in the literature on deciding how large the lower bound for the Bayes factor should be so that we're confident the evidence provided by the data surely supporting H_1 rather than H_0. Yet, since the lower bounds for the Bayes factor from the cancer data for both genders were not large enough, a tentative conclusion was that the cancer data in table 2 seemed unlikely to be misclassified. Although H_0 was not rejected for both gender in table 2 either according to their p-values (table 3, column 6), the p-value is, strictly speaking, not an appropriate measure for assessing the evidence provided by the data due to its inherent fallacy (Goodman 1999a-b).
2. From the analysis of the Bombay cancer data, the existence of Bayes factor seems to depend not only on the scenario (I or II) (the misclassification pattern), but also the multinomial distribution of p^0 (table 3). To clarify this issue, another data set related to the degree of severity for the clinical condition of myocardial infarction patients was studied (Snow 1965), where the distribution of p^0 for the treated and control groups are respectively specified as (0.4, 0.4, 0.2) and (0.3, 0.4, 0.3). It was found that the Bayes factor existed for the treated group under scenario I, but not under scenario II, whereas for the control group it exists under both scenarios. It seems that a crucial condition for the existence of Bayes factor is whether the BACST value (Eq. 13) is positive. As far as the existence of the Bayes factor is concerned, I'd like to make a conjecture which is given as follows:

"For any data set under either scenario I or II the lower bound of $\underline{B}_i^g$, i = I or II, exists if the associated $\breve{\Psi}_{K(.)}$ of Eq. 13 is positive for $K \geq 3$."

5. Conclusion

This paper addresses an issue: "how to test whether the collected categorical data are misclassified." A mixed Bayesian approach is used to test the null hypothesis that the collected data are not misclassified under a specified multinomial distribution for the studied categorical variable. The Bayes factor is employed as the main instrument to assess the evidence provided by the data. The lung cancer from all hospitals in the city of Bombay, Australia was used as an example for illustration. Based on the result of the Bayes factor in this study, the p-value was shown again not an appropriate measure to assess the evidence provided by the data.

References

Agresti, A. (2002). Categorical Data Analysis, 2nd edition. Wiley, New York.

Berger, J. O., & Selleke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and Evidence. J. Am. *Stat. Assoc., 82***,** 112-122.

Bross, I. (1954). Misclassification in 2 × 2 tables. *Biometrics, 10***,** 478-486.

Cox, D. R., & Reid, N. (1987). Approximations to noncentral distributions. Canad. J. Stat., 15**,** 105-114.

Diamond, E. L., & Lilienfeld, A. M. (1962a). Effects of errors in classification and diagnosis in various type of epidemiological studies. *Am. J. Public Health, 52***,** 1137-1144.

Diamond, E. L., & Lilienfeld, A. M. (1962b). Misclassification errors in 2 × 2 tables with one margin fixed: some further comments. *Am. J. Public Health, 52*, 2106-2110.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical Methods for Rates and Proportions, 3rd edition. Wiley, New York.

Good, I. J. (1975). The Bayes factor against equiprobability of a multinomial population assuming a symmetric Dirichlet prior. *Ann. Stat., 3*, 246-250.

Good, I. J., & Crook, J. F. (1974). The Bayes/Non-Bayes compromise and the multinomial distribution. *J. Am. Stat. Assoc., 69*, 711-720.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p value fallacy. *Ann. Intern. Med., 130*, 995-1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med., 130,* 1005-1013.

Gustafson, P. (2004). Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments. Chapman & Hall, Boca Raton, FL.

Kass, R. E., & Raftery, A. E. (1995). Bayes factor. *J. Am. Stat. Assoc., 90*, 773-795.

Lancaster, H. O. (1969). The Chi-squared Distribution. Wiley, New York.

Mote, V. L., & Anderson, R. L. (1965). An investigation of the effect of misclassification on the properties of χ2-tests in the analysis of categorical data. *Biometrika, 52*, 95-109.

Redfern, D., & Campbell, C. (1998). The MATLAB 5 Handbook. Springer, New York.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). Modern Epidemiology, 3rd edition. Lippincott Williams & Wilkins, Philadelphia, PA.

Snow, P. J. D. (1965). Effects of propranolol in myocardial infarction. *Lancet, 286*(Sept 18), 551-553.

Woodward, M. (2005). Epidemiology: Study Design and Data Analysis, 2nd ed. Chapman & Hall/CRC Press, Boca Raton, Florida.

Appendix A

With an assumption of $\theta_1 = \theta_2 = \theta_3 \equiv \theta$ and $\breve{n} = N\breve{p}$, we have under scenario I

$$\breve{n}_{(I)} \equiv \begin{bmatrix} \breve{n}_{1(I)} \\ \breve{n}_{2(I)} \\ \breve{n}_{3(I)} \end{bmatrix} = (1-3\theta)^{-1} \begin{bmatrix} (1-\theta)n_1 - \theta n_2 - \theta n_3 \\ -\theta n_1 + (1-\theta)n_2 - \theta n_3 \\ -\theta n_1 - \theta n_2 + (1-\theta)n_3 \end{bmatrix}, 0 < \theta < c_1, \tag{A1}$$

where $c_1 = \min_{j=1,2,3}\{\frac{1}{3}, \hat{p}_j\}$.

By substituting Eq. A1 into Eq. 13, we have

$$\breve{\Psi}_{3(I)} = \sum_{j=1}^{3} a_{j(I)} n_j^2 - 2\sum_{j\neq k} a_{jk(I)} n_j n_k - N, \tag{A2}$$

where

$$a_{1(I)} = (1-3\theta)^{-2}[(1-\theta)^2/n_1^0 + \theta^2(1/n_2^0 + 1/n_3^0)],$$

$$a_{2(I)} = (1-3\theta)^{-2}[(1-\theta)^2/n_2^0 + \theta^2(1/n_1^0 + 1/n_3^0)],$$

$$a_{3(I)} = (1-3\theta)^{-2}[(1-\theta)^2/n_3^0 + \theta^2(1/n_1^0 + 1/n_2^0)],$$

$$a_{12(I)} = (1-3\theta)^{-2}[\theta(1-\theta)(1/n_1^0 + 1/n_2^0) - \theta^2/n_3^0],$$

$$a_{13(I)} = (1-3\theta)^{-2}[\theta(1-\theta)(1/n_1^0 + 1/n_3^0) - \theta^2/n_2^0],$$

$$a_{23(I)} = (1-3\theta)^{-2}[\theta(1-\theta)(1/n_2^0 + 1/n_3^0) - \theta^2/n_1^0].$$

By Eq. 14, we have

$$\breve{\lambda}_3 = \sum_{j=1}^{3}(p_j^2 - 2p_j^0 p_j + p_j^{02}). \tag{A3}$$

Note that $m_g(\breve{\Psi}_K)$ of Eq. 18 with a choice of $h_0(\theta)$ which equals to the pdf of uniform distribution over [0, c_1] is reduced to

$$m_g(\breve{\Psi}_3) = \int_{\Sigma}\int_0^{c_1} \frac{1}{c_1(2+\breve{\lambda}_3)} \cdot \exp(-\frac{\breve{\Psi}_3}{2+\breve{\lambda}_3}) d\theta \cdot g(p)dp, \tag{A4}$$

where $\breve{\Psi}_{3(I)}$ and $\breve{\lambda}_3$ are given respectively by Eqs. A2 and A3. By using a linear approximation from the Taylor series expansion of $\exp(-\breve{\Psi}_3/(2+\breve{\lambda}_3))$ and another linear approximation to $(2+\breve{\lambda}_3)^{-1}$, Eq. A4 simplifies under scenario I to

$$m_g(\breve{\Psi}_{3(I)}) \approx \int_{\Sigma}\int_0^{c_1} [\tfrac{1}{4}\breve{\lambda}_3^3 - \tfrac{1}{4}(2+\breve{\Psi}_{3(I)})\breve{\lambda}_3^2 + (\breve{\Psi}_{3(I)} - 1)\breve{\lambda}_3 + 2 - \breve{\Psi}_{3(I)}] \cdot c_1^{-1} d\theta \cdot g(p)dp.$$

By substituting Eqs. A2 and A3 into the above equation and integrating $\breve{\Psi}_{3(I)}$ with respect to θ, we have after algebraic simplification

$$m_g(\dot{\Psi}_{3(I)}) = \tfrac{1}{c_1}\int_{\Sigma} g(p)\{\tfrac{1}{4}[\sum_{j=1}^{3} p_j^6 + 3\sum_{j\neq k\neq \ell} p_j^4(p_k^2 + p_\ell^2) + 6\prod_{j=1}^{3} p_j^2 - 6(\sum_{j=1}^{3} p_j^0 p_j^5 + \sum_{j\neq k\neq \ell} p_j^4(p_k^0 p_k + p_\ell^0 p_\ell))$$

$$-12(\sum_{j\neq k\neq \ell} p_j^0 p_j^3 (p_k^2 + p_\ell^2) + \sum_{j\neq k\neq \ell} p_j^0 p_j p_k^2 p_\ell^2) + 12[\sum_{j=1}^{3} p_j^{02} p_j^4 + \sum_{j\neq k} (p_j^{02} + p_k^{02}) p_j^2 p_k^2 + 2(\sum_{j\neq k} p_j^0 p_k^0 (p_j^3 p_k + p_j p_k^3)$$

$$+ \sum_{j\neq k\neq \ell} p_j^0 p_k^0 p_j p_k p_\ell^2)] + \tfrac{1}{4}(3\rho_0 - 2 - \dot{\Psi}_{3(I)}) \sum_{j=1}^{3} p_j^4 + \tfrac{1}{2}(3\rho_0 - 2 - \dot{\Psi}_{3(I)}) \sum_{j\neq k} p_j^2 p_k^2 - 2[\sum_{j=1}^{3} p_j^{03} p_j^3$$

$$+ 3 \sum_{j\neq k\neq \ell} p_j^{02} p_j^2 (p_k^0 p_k + p_\ell^0 p_\ell) + 6\prod_{j=1}^{3} p_j^0 p_j] - (3\rho_0 - 2 - \dot{\Psi}_{3(I)})[\sum_{j=1}^{3} p_j^0 p_j^3 + \sum_{j\neq k\neq \ell} p_j^0 p_j (p_k^2 + p_\ell^2)] + 3\rho_0 \sum_{j=1}^{3} p_j^{02} p_j^2$$

$$-(2 + \dot{\Psi}_{3(I)}) \sum_{j=1}^{3} p_j^0 p_j^2 + \tfrac{1}{4}[3\rho_0^2 - 4\rho_0 - 4 + 2(2 - \rho_0)\dot{\Psi}_{3(I)}] \sum_{j=1}^{3} p_j^2 + [6\rho_0 - 2(2 + \dot{\Psi}_{3(I)})] \sum_{j\neq k} p_j^0 p_k^0 p_j p_k$$

$$+ [2(\rho_0 + 1) + (\rho_0 - 2)\dot{\Psi}_{3(I)}] \sum_{j=1}^{n} p_j^0 p_j + \tfrac{1}{4}\rho_0[\rho_0^2 - 2\rho_0 - 4 + (4 - \rho_0)\dot{\Psi}_{3(I)}]\}dp\,, \tag{A5}$$

where

$$\dot{\Psi}_{3(I)} = \int_0^{c_1} \breve{\Psi}_{3(I)} d\theta = \sum_{j=1}^{3} \dot{a}_{j(I)} n_j^2 - 2\sum_{j\neq k} \dot{a}_{jk(I)} n_j n_k - N c_1\,,$$

$$\dot{a}_{1(I)} = b_{1(I)} / n_1^0 + b_{2(I)}(1/n_2^0 + 1/n_3^0)\,,$$

$$\dot{a}_{2(I)} = b_{1(I)} / n_2^0 + b_{2(I)}(1/n_1^0 + 1/n_3^0)\,,$$

$$\dot{a}_{3(I)} = b_{1(I)} / n_3^0 + b_{2(I)}(1/n_1^0 + 1/n_2^0)\,,$$

$$\dot{a}_{12(I)} = b_{3(I)}(1/n_1^0 + 1/n_2^0) - b_{2(I)} / n_3^0\,,$$

$$\dot{a}_{13(I)} = b_{3(I)}(1/n_1^0 + 1/n_3^0) - b_{2(I)} / n_2^0\,,$$

$$\dot{a}_{23(I)} = b_{3(I)}(1/n_2^0 + 1/n_3^0) - b_{2(I)} / n_1^0\,,$$

$$b_{1(I)} = \tfrac{1}{9}[c_1(5 - 3c_1)(1 - 3c_1)^{-1} - \tfrac{4}{3}\ln(1 - 3c_1)]\,,$$

$$b_{2(I)} = \tfrac{1}{9}[\tfrac{2}{3}\ln(1 - 3c_1) + c_1(2 - 3c_1)(1 - 3c_1)^{-1}]\,,$$

$$b_{3(I)} = \tfrac{1}{9}[\tfrac{1}{3}\ln(1 - 3c_1) + c_1(1 + 3c_1)(1 - 3c_1)^{-1}]\,,$$

$$\rho_0 = \sum_{j=1}^{3} p_j^{02}\,.$$

With an assumption of $\gamma_1 = \gamma_2 = \gamma_3 \equiv \gamma$, we have under scenario II

$$\breve{n}_{(II)} \equiv \begin{bmatrix} \breve{n}_{1(II)} \\ \breve{n}_{2(II)} \\ \breve{n}_{3(II)} \end{bmatrix} = [(1-\gamma)(1-3\gamma)]^{-1} \begin{bmatrix} (1-\gamma)^2 n_1 - \gamma(1-\gamma) n_2 + \gamma^2 n_3 \\ -\gamma(1-\gamma) n_1 + (1-\gamma)^2 n_2 - \gamma(1-\gamma) n_3 \\ \gamma^2 n_1 - \gamma(1-\gamma) n_2 + (1-3\gamma+\gamma^2) n_3 \end{bmatrix}, 0 < \gamma < c_2, \tag{A6}$$

where $c_2 \equiv c_1 = \min_{j=1,2,3}\{\frac{1}{3}, \hat{p}_j\}$.

By using Eq. A6, we have

$$\sum_{j=1}^{3}\breve{n}_{j(II)}^{2}=\sum_{j=1}^{3}a_{j(II)}n_j^2-2[a_{12(II)}n_1n_2+a_{23(II)}n_2n_3-a_{13(II)}n_1n_3], \tag{A7}$$

where

$$a_{1(II)}=[(1-\gamma)^4/n_1^0+\gamma^2(1-\gamma)^2/n_2^0+\gamma^4/n_3^0]/[(1-\gamma)(1-3\gamma)]^2,$$

$$a_{2(II)}=[\gamma^2(1-\gamma)^2(1/n_1^0+1/n_3^0)+(1-\gamma)^4/n_2^0]/[(1-\gamma)(1-3\gamma)]^2,$$

$$a_{3(II)}=[(\gamma^4/n_1^0+\gamma^2(1-\gamma)^2/n_2^0+(\gamma^2-3\gamma+1)^2/n_3^0]/[(1-\gamma)(1-3\gamma)]^2,$$

$$a_{12(II)}=[\gamma(1-\gamma)^3(1/n_1^0+1/n_2^0)+\gamma^3(1-\gamma)/n_3^0]/[(1-\gamma)(1-3\gamma)]^2,$$

$$a_{23(II)}=[\gamma^3(1-\gamma)/n_1^0+\gamma(1-\gamma)^3/n_2^0+\gamma(1-\gamma)(\gamma^2-3\gamma+1)/n_3^0]/[(1-\gamma)(1-3\gamma)]^2,$$

$$a_{13(II)}=[\gamma^2(1-\gamma)^2(1/n_1^0+1/n_2^0)+\gamma^2(\gamma^2-3\gamma+1)/n_3^0]/[(1-\gamma)(1-3\gamma)]^2.$$

By substituting Eq. A7 into Eq. 13 and integrating $\breve{\Psi}_{3(II)}$ with respect to γ over [0, c_2], we have

$$\dot{\Psi}_{3(II)}=\int_0^{c_2}\breve{\Psi}_{3(II)}d\gamma=\sum_{j=1}^{3}\dot{a}_{j(II)}n_j^2-2[\dot{a}_{12(II)}n_1n_2+\dot{a}_{23(II)}n_2n_3-\dot{a}_{13(II)}n_1n_3]-Nc_2, \tag{A8}$$

where

$$\dot{a}_{1(II)}=b_{1(II)}/n_1^0+b_{2(II)}/n_2^0+b_{3(II)}/n_3^0,$$

$$\dot{a}_{2(II)}=b_{1(II)}/n_2^0+b_{2(II)}(1/n_1^0+1/n_3^0),$$

$$\dot{a}_{3(II)}=b_{3(II)}/n_1^0+b_{2(II)}/n_2^0+b_{4(II)}/n_3^0,$$

$$\dot{a}_{12(II)}=b_{5(II)}(1/n_1^0+1/n_2^0)+b_{6(II)}/n_3^0,$$

$$\dot{a}_{23(II)}=b_{6(II)}/n_1^0+b_{5(II)}/n_2^0+b_{7(II)}/n_3^0,$$

$$\dot{a}_{13(II)}=b_{2(II)}(1/n_1^0+1/n_2^0)+b_{8(II)}/n_3^0,$$

$$b_{1(II)}=\int_0^{c_2}\frac{(1-\gamma)^2}{(1-3\gamma)^2}d\gamma=\tfrac{1}{27}[\frac{3c_2(5-3c_2)}{1-3c_2}-4\ln(1-3c_2)],$$

$$b_{2(II)}=\int_0^{c_2}\frac{\gamma^2}{(1-3\gamma)^2}d\gamma=\tfrac{1}{27}[\frac{3c_2(2-3c_2)}{1-3c_2}+2\ln(1-3c_2)],$$

$$b_{3(II)}=\int_0^{c_2}\frac{\gamma^4}{[(1-\gamma)(1-3\gamma)]^2}d\gamma=\tfrac{1}{108}[\frac{3c_2(12c_2^2-44c_2+14)}{(1-c_2)(1-3c_2)}+27\ln(1-c_2)+5\ln(1-3c_2)],$$

$$b_{4(II)} = \int_0^{c_2} \frac{(\gamma^2 - 3\gamma + 1)^2}{[(1-\gamma)(1-3\gamma)]^2} d\gamma = \frac{1}{108}[\frac{3c_2(12c_2^2 - 44c_2 + 14)}{(1-c_2)(1-3c_2)} - 27\ln(1-c_2) - 23\ln(1-3c_2)],$$

$$b_{5(II)} = \int_0^{c_2} \frac{\gamma(1-\gamma)}{(1-3\gamma)^2} d\gamma = \frac{1}{27}[\frac{3c_2(1+3c_2)}{1-3c_2} + \ln(1-3c_2)],$$

$$b_{6(II)} = \int_0^{c_2} \frac{\gamma^3}{(1-\gamma)(1-3\gamma)^2} d\gamma = \frac{1}{108}[\frac{3c_2(6c_2+1)}{1-3c_2} - 27\ln(1-c_2) + 7\ln(1-3c_2)],$$

$$b_{7(II)} = \int_0^{c_2} \frac{\gamma(\gamma^2 - 3\gamma + 1)}{(1-\gamma)(1-3\gamma)^2} d\gamma = \frac{1}{108}[\frac{6c_2(3c_2-2)}{1-3c_2} + 63\ln(1-c_2) - 31\ln(1-3c_2)],$$

$$b_{8(II)} = \int_0^{c_2} \frac{\gamma^2(\gamma^2 - 3\gamma + 1)}{[(1-\gamma)(1-3\gamma)]^2} d\gamma = \frac{1}{54}[\frac{3c_2(6c_2^2 + 5c_2 - 2)}{(1-c_2)(1-3c_2)} - 2\ln(1-3c_2)].$$

If the prior distribution function for g(p) is taken to be a symmetric Dirichlet's distribution with the flattening constant (or hyper-parameter) τ (τ > 0) (Good 1975), then Eq. A5 is reduced to

$$m_g(\tau \mid \dot{\Psi}_{3(I)}) = \frac{1}{c_1} \frac{d_4\tau^4 + d_3\tau^3 + d_2\tau^2 + d_1\tau + d_0}{12(3\tau+1)(3\tau+2)(3\tau+4)(3\tau+5)}, \quad \text{(A9)}$$

where

$$d_4 = 243\rho_0^3 - 1377\rho_0^2 + 891\rho_0 + (648\rho_0 - 432)\rho_1 - 216\rho_2 - 72\rho_3 - (243\rho_0^2 - 1782\rho_0 + 432\rho_1 + 1647)\dot{\Psi}_{3(I)} + 1059,$$

$$d_3 = 972\rho_0^3 - 5130\rho_0^2 + 2568\rho_0 + (1176\rho_0 + 464)\rho_1 - 864\rho_2 - 432\rho_3 - (972\rho_0^2 - 7020\rho_0 + 392\rho_1 + 6470)\dot{\Psi}_{3(I)} + 4382,$$

$$d_2 = 1323\rho_0^3 - 6101\rho_0^2 + 5307\rho_0 + (2736\rho_0 + 1728)\rho_1 - 1128\rho_2 - 1240\rho_3 - (1323\rho_0^2 - 9306\rho_0 + 912\rho_1 + 8425)\dot{\Psi}_{3(I)} + 5943,$$

$$d_1 = 702\rho_0^3 + 2450\rho_0^2 + 138\rho_0 + (960\rho_0 + 1920)\rho_1 - 1440\rho_2 - 2736\rho_3 - (702\rho_0^2 - 4692\rho_0 + 320\rho_1 - 5274)\dot{\Psi}_{3(I)} + 2722,$$

$$d_0 = 40[3\rho_0^3 - 15\rho_0^2 + 48\rho_0 - 24\rho_3 - (3\rho_0^2 - 18\rho_0 - 14)\dot{\Psi}_{3(I)} + 3],$$

$$\rho_1 = \sum_{j \neq k} p_j^0 p_k^0,$$

$$\rho_2 = \sum_{j \neq k \neq \ell} (p_j^0 + p_k^0) p_\ell^{02},$$

$$\rho_3 = \sum_{j=1}^{3} p_j^{03}.$$

Similarly, $m_g(\tau \mid \dot{\Psi}_{3(II)})$ has exactly the same expression like Eq. A9 except that $\dot{\Psi}_{3(I)}$ and c_1 are replaced respectively by $\dot{\Psi}_{3(II)}$ and c_2.

To avoid the use of hyper-prior distribution on τ (Good and Crook 1974), the non-Bayesian approach is used to find the stationary point $\tau_{\max(.)}$ for $m_g(\tau \mid \breve{\Psi}_{3(.)})$. By using an elementary technique in calculus to calculate the first derivative

of $m_g(\tau \mid \breve{\Psi}_{3(I)})$ and set it equal to zero, we have after simplification

$$(324d_4 - 81d_3)\tau^6 + (882d_4 - 162d_2)\tau^5 + (702d_4 + 441d_3 - 324d_2 - 243d_1)\tau^4 + (160d_4 + 468d_3 - 648d_1 - 324d_0)\tau^3$$

$$+(120d_3 + 234d_2 - 441d_1 - 972d_0)\tau^2 + (80d_2 - 882d_0)\tau + 40d_1 - 234d_0 = 0 . \quad \text{(A10)}$$

To solve Eq. A10 for the stationary points, I employed the "ROOTS" subroutine in the MATLAB (Redfern and Campbell 1998).

According to the terms of Good (1975), the way to estimate $\tau_{\max}$ is called by the type II maximum likelihood or the maximum hyper-prior likelihood method. This kind of approach to estimate the Bayes factor is called the Bayesian/Fisherian criterion which is a compromise from taking a full Bayesian approach.

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

On Comparison of Local Polynomial Regression Estimators for $P = 0$ and $P = 1$ in a Model Based Framework

Conlet Biketi Kikechi[1] & Richard Onyino Simwa[2]

[1] Statistics and Operations Research Section, School of Mathematics, College of Biological and Physical Sciences, University of Nairobi, Nairobi, Kenya

[2] Actuarial Science and Financial Mathematics Section, School of Mathematics, College of Biological and Physical Sciences, University of Nairobi, Nairobi, Kenya

Correspondence: Conlet Biketi Kikechi, Statistics and Operations Research Section, School of Mathematics, College of Biological and Physical Sciences, University of Nairobi, Nairobi, Kenya. Email: Kikechiconlet@gmail.com

Received: May 16, 2018 Accepted: May 31, 2018 Online Published: June 28, 2018

doi:10.5539/ijsp.v7n4p104 URL: https://doi.org/10.5539/ijsp.v7n4p104

Abstract

This article discusses the local polynomial regression estimator for $P = 0$ and the local polynomial regression estimator for $P = 1$ in a finite population. The performance criterion exploited in this study focuses on the efficiency of the finite population total estimators. Further, the discussion explores analytical comparisons between the two estimators with respect to asymptotic relative efficiency. In particular, asymptotic properties of the local polynomial regression estimator of finite population total for $P = 0$ are derived in a model based framework. The results of the local polynomial regression estimator for $P = 0$ are compared with those of the local polynomial regression estimator for $P = 1$ studied by Kikechi et al (2018). Variance comparisons are made using the local polynomial regression estimator $\bar{T}_0$ for $\mathrm{P} = 0$ and the local polynomial regression estimator $\bar{T}_1$ for $\mathrm{P} = 1$ which indicate that the estimators are asymptotically equivalently efficient. Simulation experiments carried out show that the local polynomial regression estimator $\bar{T}_1$ outperforms the local polynomial regression estimator $\bar{T}_0$ in the linear, quadratic and bump populations.

Keywords: Asymptotic Properties, Asymptotic Relative Efficiency, Finite Population, Local Polynomial Regression, Model Based Framework, Nonparametric Regression, Sample Surveys

1. Introduction

The theory of sample surveys involves principles and methods of collecting and analyzing data from a finite population of N units and then making inferences about finite population parameters on the basis of information obtained from the sample. For some early work on survey sampling theory, see Royall (1970a), Royall (1970b), Royall (1971), Smith (1976) and Pfeffermann (1993). In this study, an estimator of the finite population total is developed and its properties derived using the local polynomial regression procedure. Local polynomial regression is a nonparametric technique which is a generalization of kernel regression and is used for smoothing scatter plots and modeling functions. Under normal conditions, when $p = 0$, this is referred to as local constant regression, when $p = 1$, this is local linear regression and when $p \geq 2$, this is local polynomial regression. p is the order of the local polynomial being fit. In local polynomial regression, a low order weighted least squares regression is fit at each point of interest x, using data from some neighborhood around x (see Cleveland (1979) and Cleveland and Devlin (1988)).

Once a modeling approach is undertaken, there is a special feature in finite population estimation problems that the unknown quantities are realized values of random variables, so the basic problem has the feature of being similar to a prediction problem. In order to estimate $m(x)$ at a given point x, the association between the predictor variable and the response variable is explored. This methodology was introduced by Stone (1977). It has also been studied by Fan (1993), Fan and Gijbels (1996), Breidt and Opsomer (2000) and Kikechi et al (2017). Like in Stone (1977), the main aim of this procedure is to quantify the contribution of the covariate X to the response Y per unit value of X in order to summarize the association between the two variables, to predict the mean response for a given value X and to extrapolate the results beyond the range of the observed covariate values. A weight $k\left(\frac{x_i - x}{h}\right)$ is assigned to the point

(x_i, y_i) where h is the size of the local neighbourhood and $k(t)$ is the unimodal non-negative function. On the other hand, inferences may explore properties of the process that generate the population values (Montanari and Ranalli (2003)). An assumption is made from the fact that the finite population has been generated by a super population model $\xi = f(x, y, \varphi)$ and it is of interest to estimate the population parameters φ, where $\varphi = \alpha + \beta x_i$. The super population model can be applied to predict the unobserved values $y_i's$ after obtaining estimates of α and β using the known auxiliary information $x_i,\ i = 1,2 \dots, N$ (see Montanari and Ranalli (2005) and Rueda and Sanchez-Borrego (2009)).

The nonparametric approach does not restrict the functional form of the distribution nor does it specify the various stochastic properties such as $E_\xi(.)$, $V_\xi(.)$ and $MSE_\xi(.)$. Rather, it leaves them to cover broad classes of models, thus allowing for more robust inference than inference obtained in parametric approach. Using the model ξ, the nonparametric estimator of total, T has been derived by Nadaraya (1964), Watson (1964), Priestly and Chao (1972), Gasser and Muller (1979), Dorfman (1992)), Chambers et al (1993) and Odhiambo and Mwalili (2000). In his study, Dorfman (1992) has been able to prove the asymptotic unbiasedness and MSE consistency of this estimator. The estimator, however suffers from sparse sample problem, and more work needs to be done to come up with another technique that can overcome this problem. This is where the local polynomial procedure comes in. See Kikechi et al (2017) and Kikechi et al (2018).

The local polynomial regression is one of the most successfully applied design adaptive non parametric regression. This estimation procedure is an attractive choice due to its flexibility and asymptotic performance. Having a local model (rather than just a point estimate) enables derivation of response adaptive methods for bandwidth and polynomial order selection in a straightforward manner. The procedure has also the advantage of eliminating design bias and alleviating boundary bias. Furthermore, the method adapts well to random, fixed, highly clustered and nearly uniform designs. The weighted least squares principle to be employed in the local polynomial approximation approach, opens the way to a wealth of statistical knowledge and thus providing easy computations and generalizations. See Fan (1992), Fan (1993), Ruppert and Wand (1994) and Fan and Gijbels (1996) among others.

Kikechi et al (2018) employ a superpopulation approach to estimate the finite population total using the procedure of local linear regression. Explicitly, the authors derive robustness properties of the local linear regression estimator and carry out simulation experiments on the performances of this estimator in comparison with other estimators that exist in the literature. Results indicate that the local linear regression estimator is more efficient and performing better than the Horvitz-Thompson (1952) and Dorfman (1992) estimators, regardless of whether the model is specified or mispecified. In this paper, the local polynomial regression estimator of finite population total for $P = 0$ is studied and asymptotic properties derived. Analytical comparisons are carried out between this estimator and the local polynomial regression estimator for $P = 1$ studied by Kikechi et al (2018) which indicate that the estimators are asymptotically equivalently efficient. Simulation experiments however indicate that the local polynomial regression estimator $\bar{T}_1$ is superior and dominates the local polynomial regression estimator $\bar{T}_0$ in the linear, quadratic and bump populations.

2. Method of Constructing the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

The superpopulation model considered for estimating the finite population total is given by,

$$Y_i = m(X_i) + \sigma^2(X_i)\varepsilon_i \tag{1}$$

Specifically, the following assumptions hold for the model considered in the nonparametric regression estimation of $m(x_i)$:

$$E(Y_i/X_i = x_i) = m(x_i)$$

$$Cov\left(Y_i, Y_j/X_i = x_i, X_j = x_j\right) = \begin{cases} \sigma^2(x_i), & i = j \\ 0 \quad , & i \neq j \end{cases} \quad i = 1, 2, 3, \dots, N \quad j = 1,2,3, \dots, N\,. \tag{2}$$

The properties of the error are given by,

$$E(\varepsilon_i/X_i = x_i) = m(x_i)$$

$$Cov\left(\varepsilon_i, \varepsilon_j/X_i = x_i, X_j = x_j\right) = \begin{cases} \sigma^2(x_i), & i = j \\ 0 \quad , & i \neq j \end{cases} \quad i = 1, 2, 3, \dots, N \quad j = 1,2,3, \dots, N\,. \tag{3}$$

The functions $m(x_i)$ and $\sigma^2(x_i)$ are assumed to be smooth and strictly positive. Consider the Taylor series

expansion of $m(x_i)$ expressed as,

$$m(x_i) = m(x_j + ht) = m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!}m''(x_j) + \frac{h^3t^3}{3!}m'''(x_j) + \cdots$$
$$= m(x_j) + (x_i - x_j)m'(x_j) + \frac{(x_i - x_j)^2}{2!}m''(x_j) + \frac{(x_i - x_j)^3}{3!}m'''(x_j) + \cdots \quad (4)$$

The Taylor series expansion is written in a general form expressed as,

$$y_i = \alpha + (x_i - x_j)\beta + \varepsilon_i \quad (5)$$

where x_i lies in the interval $[x_j - h,\ x_j + h]$ and

$$\varepsilon_i = \frac{(x_i - x_j)^2}{2!}m''(x_j) + \frac{(x_i - x_j)^3}{3!}m'''(x_j) + \cdots$$

The constants α and β are solved using the least squares procedure by making ε_i the subject of the formulae, squaring both sides, summing over all possible sample values and applying the weights to obtain a solution to the weighted least squares problem of the form;

$$\sum_{i\in S}\varepsilon_i^2 = \sum_{i\in S}\left(y_i - \alpha - \beta(x_i - x_j)\right)^2 K\left(\frac{x_i - x_j}{h}\right) \quad (6)$$

Letting,

$$\varphi = \sum_{i\in S}\left(y_i - \alpha - \beta(x_i - x_j)\right)^2 K\left(\frac{x_i - x_j}{h}\right) \quad (7)$$

Differentiating φ with respect to α and equating to zero, gives

$$\frac{\partial\varphi}{\partial\alpha} = \sum_{i\in S} -2\left(y_i - \alpha - \beta(x_i - x_j)\right)K\left(\frac{x_i - x_j}{h}\right)\left\{\left\{\sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right)\right\}^{-1}\right\} = 0 \quad (8)$$

Implying that

$$\sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right)y_i = \alpha\sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right) + \beta\sum_{i\in S}(x_i - x_j)K\left(\frac{x_i - x_j}{h}\right). \quad (9)$$

Letting

$$S_{n,l} = \sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)^l \quad (10)$$

Then it follows from equation (9) that

$$\sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right)y_i = \alpha(S_{n,0}) + \beta(S_{n,1}). \quad (11)$$

Similarly, differentiating φ with respect to β and equating to zero, gives

$$\frac{\partial\varphi}{\partial\beta} = \sum_{i\in S} -2\left(y_i - \alpha - \beta(x_i - x_j)\right)(x_i - x_j)K\left(\frac{x_i - x_j}{h}\right)\left\{\left\{\sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right)\right\}^{-1}\right\} = 0 \quad (12)$$

Implying that

$$\sum_{i\in S}(x_i - x_j)K\left(\frac{x_i - x_j}{h}\right)y_i = \alpha\sum_{i\in S}(x_i - x_j)K\left(\frac{x_i - x_j}{h}\right) + \beta\sum_{i\in S}(x_i - x_j)^2K\left(\frac{x_i - x_j}{h}\right). \quad (13)$$

and thus

$$\sum_{i\in S}(x_i - x_j)K\left(\frac{x_i - x_j}{h}\right)y_i = \alpha(S_{n,1}) + \beta(S_{n,2}). \quad (14)$$

Multiplying equation (11) and equation (14) by $(S_{n,2})$ and $(S_{n,1})$ respectively, gives

$$(S_{n,2})\sum_{i\in S} K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha(S_{n,0})(S_{n,2}) + \beta(S_{n,1})(S_{n,2}) \tag{15}$$

$$(S_{n,1})\sum_{i\in S} (x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha(S_{n,1})^2 + \beta(S_{n,1})(S_{n,2}) \tag{16}$$

Subtracting equation (16) from equation (15), gives

$$(S_{n,2})\sum_{i\in S} K\left(\frac{x_i - x_j}{h}\right) y_i - (S_{n,1})\sum_{i\in S} (x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha(S_{n,0})(S_{n,2}) - \alpha(S_{n,1})^2 \tag{17}$$

Making α the subject of the formulae, gives

$$\bar{\alpha} = \sum_{i\in S}\left\{\frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i\right\} \tag{18}$$

Similarly, multiplying equation (11) and equation (14) by $(S_{n,1})$ and $(S_{n,0})$ respectively, gives

$$(S_{n,1})\sum_{i\in S} K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha(S_{n,0})(S_{n,1}) + \beta(S_{n,1})^2 \tag{19}$$

$$(S_{n,0})\sum_{i\in S} (x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha(S_{n,0})(S_{n,1}) + \beta(S_{n,0})(S_{n,2}) \tag{20}$$

Subtracting equation (20) from equation (19), gives

$$(S_{n,1})\sum_{i\in S} K\left(\frac{x_i - x_j}{h}\right) y_i - (S_{n,0})\sum_{i\in S} (x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) y_i = \beta(S_{n,1})^2 - \beta(S_{n,0})(S_{n,2}) \tag{21}$$

Making β the subject of the formulae, gives

$$\bar{\beta} = \sum_{i\in S}\left\{\frac{(S_{n,0}(x_i - x_j) - S_{n,1})}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i\right\} \tag{22}$$

Now it follows from equation (5) that

$$\bar{y}_i = \bar{\alpha} + (x_i - x_j)\bar{\beta} \tag{23}$$

If the value assigned is zero, assuming that $\bar{\beta}$ is a pre-assigned constant, then

$$\bar{y}_j = \bar{\alpha} \tag{24}$$

Therefore

$$\bar{m}(x_j) = \sum_{i\in S}\left\{\frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i\right\}$$

$$= \sum_{i\in S} w_i(x_j) y_i \tag{25}$$

where

$$w_i(x_j) = \frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i$$

Implying that the finite population total estimator $\bar{T}$ for $P = 0$ can be estimated using

$$\bar{T} = \sum_{i\in S} y_i + \sum_{j\in R} \bar{m}(x_j)$$

$$= \sum_{i\in S} y_i + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i\right\}\right\} \tag{26}$$

3. Properties of the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

In deriving the properties of the local polynomial regression estimator, the following assumptions are made according to Ruppert and Wand (1994):

(i) The x_j variables lie in the interval $(0, 1)$.

(ii) The function $m''(.)$ is bounded and continuous on $(0, 1)$.

(iii) The kernel $K(t)$ is symmetric and supported on $(-1, 1)$. Also $K(t)$ is bounded and continuous satisfying the following: $\int_{-\infty}^{\infty} K(x)\,dx = 1$, $\int_{-\infty}^{\infty} xK(x)\,dx = 0$, $\int_{-\infty}^{\infty} x^2K(x)\,dx > 0$, $\int_{-\infty}^{\infty} K^2x\,dx < \infty$, $d_k = \int_{-\infty}^{\infty} K^2(t)\,dt$

(iv) The bandwidth h is a sequence of values which depend on the sample size n and satisfying $h \to 0$ and $nh \to \infty$, as $n \to \infty$.

(v) The point x_j at which the estimation is taking place satisfies $h < x_j < 1 - h$.

Fan (1993) imposed conditions on $K(.)$ and are only used for convenience in terms of technical arguments and thus can be relaxed.

3.1 The Expectation of the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

The expectation of $\bar{T}$ *for* $P = 0$ is derived as,

$$E(\bar{T}) = \sum_{i\in S} E(y_i) + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} k\left(\frac{x_i - x_j}{h}\right) E(y_i)\right\}\right\}$$
$$= \sum_{i\in S} m(x_i) + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{S_{n,0}S_{n,2} - (S_{n,1})^2} k\left(\frac{x_i - x_j}{h}\right) m(x_i)\right\}\right\} \quad (27)$$

Using the Taylor series expansion of the form,

$$m(x_i) = m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!}m''(x_j) + \cdots, \quad (28)$$

Theorem 3 in Fan and Gijbels (1996) is such that under the conditions given in (i)-(v), allows

$$E(\bar{T}) = \sum_{i\in S} m(x_i) + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{S_{n,2} k\left(\frac{x_i - x_j}{h}\right)}{S_{n,0}S_{n,2} - (S_{n,1})^2}\left(m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!}m''(x_j) + \cdots\right)\right\}\right\}$$
$$- \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{S_{n,1}(x_i - x_j)}{S_{n,0}S_{n,2} - (S_{n,1})^2} k\left(\frac{x_i - x_j}{h}\right)\left(m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!}m''(x_j) + \cdots\right)\right\}\right\}$$
$$= \sum_{i\in S} m(x_i) + \sum_{j\in R}\left\{\left(\frac{S_{n,0}S_{n,2} - (S_{n,1})^2}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right) m(x_j)\right\} + \sum_{j\in R}\left\{\left(\frac{S_{n,1}S_{n,2} - S_{n,1}S_{n,2}}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right) m'(x_j)\right\}$$
$$+ \sum_{j\in R}\left\{\left(\frac{(S_{n,2})^2 - S_{n,1}S_{n,3}}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)\frac{m''(x_j)}{2}\right\}$$
$$= \sum_{i\in S} m(x_i) + \sum_{j\in R} m(x_j) + \sum_{j\in R}\left\{\left(\frac{(S_{n,2})^2 - S_{n,1}S_{n,3}}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)\frac{m''(x_j)}{2}\right\}. \quad (29)$$

3.2 The Bias of the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

The bias of $\bar{T}$ is given by

$$Bias(\bar{T}) = \sum_{j\in R}\left\{\left(\frac{(S_{n,2})^2 - S_{n,1}S_{n,3}}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)\frac{m''(x_j)}{2}\right\}. \quad (30)$$

Therefore the asymptotic expression of the bias of the local polynomial regression estimator $\bar{T}$ is

$$Bias_{asy}(\bar{T}) = \sum_{j\in R}\left\{\frac{\left(n^2h^6{k_2}^2 + o(n^2h^8)\right)m''(x_j)}{2\left(n^2h^4k_2 + o(n^2h^6)\right)}\right\}$$
$$= \sum_{J\in R}\left\{\frac{1}{2}h^2k_2m''(x_j)\right\} \quad (31)$$

3.3 The Variance of the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

The variance of the local polynomial regression estimator $\bar{T}$ is estimated using the variance of the error, thus $Var(\bar{T} - T)$ is derived as

$$\begin{aligned} Var(\bar{T}) &= Var\left\{\sum_{i\in S} y_i + \sum_{j\in R} \bar{m}(x_j) - \sum_{i\in S} y_i - \sum_{j\in R} y_j\right\} \\ &= Var\left\{\sum_{i\in S}\sum_{j\in R} w_i(x_j) y_i - \sum_{j\in R} y_j\right\} \\ &= \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \end{aligned} \tag{32}$$

where,

$$w_i(x_j) = \frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right).$$

The asymptotic expression for the variance of $\bar{T}$ is given by the expression using the results of $\bar{m}(x_j)$ that have been derived, thus

$$\begin{aligned} Var_{asy}(\bar{T}) &= \frac{1}{nh}\sum_{j\in R}\sum_{i\in S}\left\{K^2\left(\frac{x_i - x_j}{h}\right)\sigma^2(x_i)\left(\frac{x_i - x_{i-1}}{h}\right)\right\} \\ &= \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j). \end{aligned} \tag{33}$$

3.4 The MSE of the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

Theorem I in Fan (1993) allows that under condition (ii) gives,

$$\begin{aligned} MSE(\bar{T}) &= \{Bias(\bar{T})\}^2 + Var(\bar{T}) \\ &= \left\{\sum_{j\in R}\left\{\left(\frac{(S_{n,2})^2 - S_{n,1}S_{n,3}}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)\frac{m''(x_j)}{2}\right\}\right\}^2 + \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \end{aligned} \tag{34}$$

The asymptotic expression for the MSE of the local polynomial regression estimator $\bar{T}$ is given by

$$MSE_{asy}(\bar{T}) = \left\{\sum_{j\in R}\left\{\frac{1}{2}\,h^2 k_2 m''(x_j)\right\}\right\}^2 \tag{35}$$

Note that results for the local polynomial regression estimator of finite population total $\bar{\mathrm{T}}$ for $\mathrm{P} = 1$ have been derived by Kikechi et al (2018).

3.5 The Asymptotic Relative Efficiency

The relative efficiency of two procedures is the ratio of their efficiencies, but it is often possible to use the asymptotic relative efficiency, defined as the limit of the relative efficiencies as the sample size grows, as the principal measure of comparison. Let $\bar{T}_0$ be the local polynomial regression estimator of finite population total for $\mathrm{P} = 0$ and $\bar{T}_1$ be the local polynomial regression estimator of finite population total for $\mathrm{P} = 1$ as studied by Kikechi et al (2018).

If $\bar{T}_0$ and $\bar{T}_1$ are both unbiased estimators of T, then the relative efficiency of $\bar{T}_0$ to $\bar{T}_1$ is given by,

$$Eff(\bar{T}_0, \bar{T}_1) = \frac{Var(\bar{T}_1)}{Var(\bar{T}_0)}. \tag{36}$$

If $\bar{T}_0$ and $\bar{T}_1$ are both asymptotically unbiased estimators of T, then the asymptotic relative efficiency of $\bar{T}_0$ to $\bar{T}_1$ is given by,

$$ARE(\bar{T}_0, \bar{T}_1) = \lim_{n\to\infty} Eff(\bar{T}_0, \bar{T}_1) = \lim_{n\to\infty}\frac{Var(\bar{T}_1)}{Var(\bar{T}_0)}. \tag{37}$$

Therefore, the estimators of finite population totals for $\bar{T}_0$ and $\bar{T}_1$ are respectively given by,

$$\bar{T}_0 = \sum_{i\in S} y_i + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i\right\}\right\}. \tag{38}$$

$$\bar{T}_1 = \sum_{i\in S} \mathrm{Y}_i + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} k\left(\frac{x_i - x_j}{h}\right) y_i\right\}\right\}$$
$$+ \sum_{\boldsymbol{j\in R}}\left\{\left(\frac{x_i - x_j}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)\sum_{\boldsymbol{i\in S}}\left\{(S_{n,0}(x_i - x_j) - S_{n,1})k\left(\frac{x_i - x_j}{h}\right) y_i\right\}\right\}. \tag{39}$$

The variance of the local polynomial regression estimator $\bar{T}_0$ is given by,

$$Var(\bar{T}_0) \quad = \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \tag{40}$$

The asymptotic expression for the variance of the local polynomial regression estimator $\bar{T}_0$ is estimated by,

$$Var_{asy}(\bar{T}_0) = \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j) \tag{41}$$

The variance of the local polynomial regression estimator $\bar{T}_1$ is given by,

$$Var(\bar{T}_1\ ) = \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}(x_i - x_j)^2\sum_{i\in S} w_i'^2(x_j)\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \tag{42}$$

The asymptotic expression for the variance of the local polynomial regression estimator $\bar{T}_1$ is estimated by,

$$Var_{asy}(\bar{T}_1) = \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j)\,. \tag{43}$$

Note that in Kikechi e tal (2017), $Var_{asy}\left(\bar{m}_{LL}(x_j)\right) = \frac{d_k}{nh}\sigma^2(x_j)$ and $Var_{asy}\left(\bar{m}_{NW}(x_j)\right) = \frac{d_k}{nh}\sigma^2(x_j)$

Thus the asymptotic relative efficiency of the local polynomial regression estimator $\bar{T}_0$ to the local polynomial regression estimator $\bar{T}_1$ derived by Kikechi et al (2018) is given by,

$$ARE(\bar{T}_0, \bar{T}_1) = \lim_{n\to\infty} Eff(\bar{T}_0, \bar{T}_1) = \lim_{n\to\infty}\left\{\frac{Var_{asy}(\bar{T}_1)}{Var_{asy}(\bar{T}_0)}\right\} = \lim_{n\to\infty}\left\{\frac{\sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j)}{\sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j)}\right\} = 1\,. \tag{44}$$

4. Simulation Study

4.1 Description of the Data Sets

In this section, simulation experiments are carried out to evaluate the performance of the estimators. The data are generated from the regression model of the form,

$$Y_i = m(X_i) + \sigma^2(X_i)\varepsilon_i \qquad i = 1,2,\dots,n \tag{45}$$

The data sets are obtained by simulation using specific models having relations of the form,

$$y_i = 1 + 2(x - 0.5) + \varepsilon_i \tag{46}$$

$$y_i = 1 + 2(x - 0.5)^2 + \varepsilon_i \tag{47}$$

$$y_i = 1 + 2(x - 0.5) + \exp(-200(x - 0.5)^2 + \varepsilon_i \tag{48}$$

for the linear, quadratic and bump populations respectively. The $x_i's$ are generated as independent and identically distributed (iid) uniform (0, 1) random variables. The errors are assumed to be independent and identically distributed (iid) random variables with mean 0 and constant variance. The analysis and comparison in terms of performance is based on the local polynomial regression estimator $\bar{T}_0$ and the local polynomial regression estimator $\bar{T}_1$. The Epanechnicov kernel given is used for kernel smoothing on each of the populations due to its simplicity and easy computations using well designed computer programs and is defined as,

$$\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}t^2\right)|t|<\sqrt{5} \tag{49}$$

The bandwidths are data driven and are determined by the least squares cross validation method. For each of the three artificial populations of size 200, samples are generated by simple random sampling without replacement using sample size $n=60$. For each combination of mean function, standard deviation and bandwidth, 500 replicate samples are selected and the estimators calculated.

Table 1. Computational Formulae for the Local Polynomial Regression Estimators $\bar{T}_0$ and $\bar{T}_1$

Estimator	Formulae
$LPRE,\bar{T}_0$	$\bar{T}_0=\sum_{i\in S}Y_i+\sum_{j\in R}\bar{m}_0\left(x_j\right)$
$LPRE,\bar{T}_1$	$\bar{T}_1=\sum_{i\in S}Y_i+\sum_{j\in R}\bar{m}_1\left(x_j\right)$

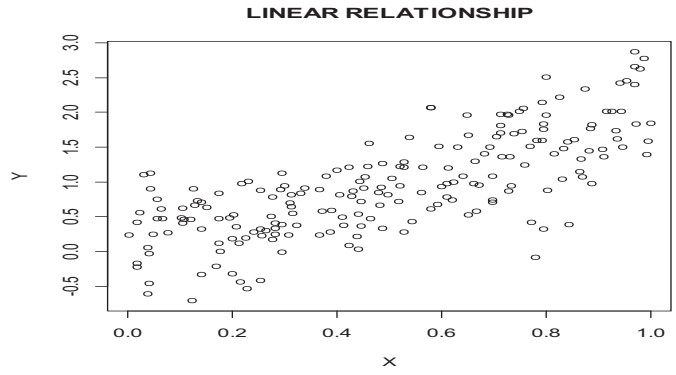


Figure 1. Scatter Diagram for the Linear Population

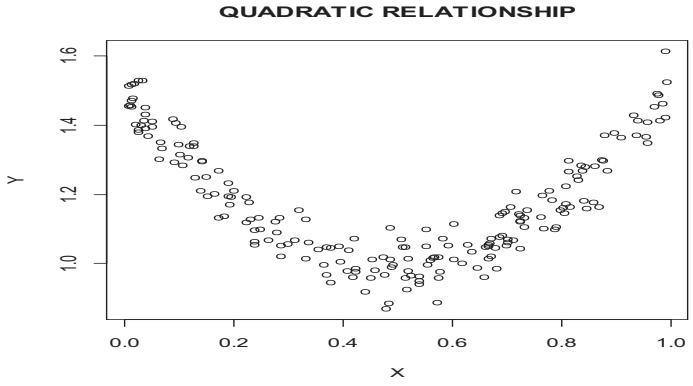


Figure 2. Scatter Diagram for the Quadratic Population

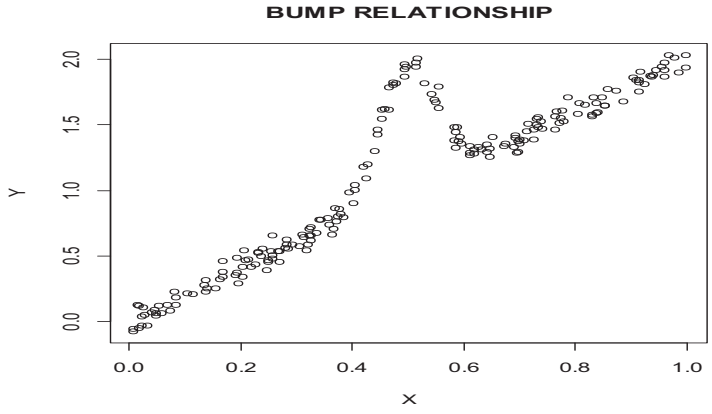


Figure 3. Scatter Diagram for the Bump Population

4.2 Results

The results of the bias and mean squared error (MSE) for the local polynomial regression estimator $\bar{T}_0$ for $P = 0$ and the local polynomial regression estimator $\bar{T}_1$ for $P = 1$ in the linear, quadratic and bump populations are provided in the table below.

Table 2. The Bias and MSE for $\bar{T}_0$ and $\bar{T}_1$ in the Three Artificial Populations

	Linear		**Quadratic**		**Bump**	
	$\bar{T}_0$	$\bar{T}_1$	$\bar{T}_0$	$\bar{T}_1$	$\bar{T}_0$	$\bar{T}_1$
BIAS	5.507608	3.777348	4.7372	0.45116	5.293896	0.4187236
MSE	100.8874	15.40735	18.40769	0.1601695	43.9272	0.1896261

5. Discussion

In estimating $\bar{m}(x_j)$ for the local polynomial regression estimator $\bar{T}_0$, $\bar{\beta}$ has been assumed to be a pre-assigned constant and in particular the value assigned is zero. It has therefore been shown in section 2 that the estimator $\bar{m}(x_j)$ is biased leading to a biased estimation of the finite population total. On the other hand, when estimating $\bar{m}(x_j)$ for the local polynomial regression estimator $\bar{T}_1$, the value of $\bar{\beta}$ is not pre-assigned but rather determined by the set of data provided and thus minimizing the bias. With regard to asymptotic relative efficiency, there is no difference in the performance of the local polynomial regression estimator $\bar{T}_0$ studied in this paper and the local polynomial regression estimator $\bar{T}_1$ studied by Kikechi et al (2018). The reason for this being that their ratio converges to 1 as n becomes large, see equation (44). This therefore implies that the estimators are asymptotically equivalently efficient. However, it is observed from simulation experiments conducted that the biases and MSEs computed in table 2 for the local polynomial regression estimator $\bar{T}_1$ are small in all the three populations. The results therefore indicate that the local polynomial regression estimator $\bar{T}_1$ is superior and dominates the local polynomial regression estimator $\bar{T}_0$ for the linear, quadratic and bump populations.

6. Conclusion

In this article the local polynomial regression estimators $\bar{T}_0$ and $\bar{T}_1$ of finite population totals have been studied in a model based framework. Analytically, variance comparisons are explored using the local polynomial regression estimator $\bar{T}_0$ for $\mathrm{P} = 0$ and the local polynomial regression estimator $\bar{T}_1$ for $\mathrm{P} = 1$ in which results indicate that the estimators are asymptotically equivalently efficient. Simulation experiments carried out in terms of the biases and MSEs show that the local polynomial regression estimator $\bar{T}_1$ outperforms the local polynomial regression estimator $\bar{T}_0$ in all the three artificial populations and therefore, $\bar{T}_1$ is the most efficient estimator.

References

Breidt, F. J., & Opsomer, J. D. (2000). Local Polynomial Regression Estimation in Survey Sampling. *Annals of statistics, 28*, 1026-1053.

Chambers, R. L., Dorfman, A. H., & Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *J. Amer Statist Assoc., 88,* 268-277.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatter Plots. *J. Amer. Statist. Assoc. 74*, 829–836.

Cleveland, W. S., & Devlin, S. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J. Amer. Statist. Assoc. 83*, 596–610.

Dorfman, A. (1992). Nonparametric Regression for Estimating Totals in Finite Populations, Proceedings of the Section on Survey Research Methods. *American Statistical Association*, 622-625.

Fan, J. (1992). Design Adaptive Nonparametric Regression. *Journal of American Statistical Association, 87*, 998-1004.

Fan, J. (1993). Local Linear Regression Smoothers and Their Minimax Efficiencies. *Annals of Statistics, 21*, 196–216. https://doi.org/10.1214/aos/1176349022

Fan, J., & Gijbels, I. (1996). Local Polynomial Modeling and its Applications. London: Chapman and Hall.

Gasser, T., & Muller, H. G. (1979). Kernel Estimation in Regression Functions. *Smoothing Techniques for Curve Estimation*, 23-68.

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of American Statistical Association, 47*, 663-685. https://doi.org/10.1080/01621459.1952.10483446

Kikechi, C. B., Simwa, R. O., & Pokhariyal, G. P. (2017). On Local Linear Regression Estimation in Sampling Surveys. *Far East Journal of Theoretical Statistics, 53*(5), 291-311. . https://doi.org/10.17654/TS053050291

Kikechi, C. B., Simwa, R. O., & Pokhariyal, G. P. (2018). On Local Linear Regression Estimation of Finite Population Totals in Model Based Surveys. *American Journal of Theoretical and Applied Statistics, 7*(3), 92-101. . https://doi.org/10.11648/j.ajtas.20180703.11

Montanari, G. E., & Ranalli, M. G. (2003). Nonparametric Methods in Survey Sampling. In: Vinci, M., Monari, P., Mignani, S. and Montanari, A., Eds., New Developments in Classification and Data Analysis, Springer, Berlin, 203-210.

Montanari, G. E., & Ranalli, M. G. (2005). Nonparametric Model Calibration Estimation in Survey Sampling. *Journal of the American Statistical Association, 100*, 1429-1442. https://doi.org/10.1198/016214505000000141

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability Applications, 10*, 186-190.

Odhiambo, R. O., & Mwalili, T. (2000). Nonparametric Regression for Finite Population Estimation. *East African Journal of Science, II*(2), 107-112.

Pfeffermann, D. (1993). The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review, 61*(2), 317-337. https://doi.org/10.2307/1403631

Priestley, M. B., & Chao, M. T. (1972). Nonparametric Function Fitting. *Journal of the Royal Statistical Society, B34*, 384-392.

Royall, R. M. (1970a). On Finite Population Sampling under certain Linear Regression Models. *Biometrika, 57,* 377-387

Royall, R. M. (1970b). Finite Population Sampling-On Labels in Estimation. *Journal of the Annals of Mathematical Statistics, 41*, 1774-1779.

Royall, R. M. (1971). Linear Regression Models in Finite Population Sampling Theory Holt, Rinhart and Winston, Toronto, Canada, 54, 499-513.

Rueda, M. & Sanchez-Borrego, I. (2009). A Predictive Estimator of Finite Population Mean Using Nonparametric Regression. *Computational Statistics 24*, 1-14. https://doi.org/10.1007/s00180-008-0140-x

Ruppert, D., & Wand, M. P. (1994). Multivariate Locally Weighted Least Squares Regression. *Annals of Statistics, 22*, 1346-1370. https://doi.org/10.1214/aos/1176325632

Smith, T. M. (1976). The Foundations of Survey Sampling. *Journal of Royal Statistical Society Association, 139*, Part 2 183-204.

Stone, C. (1977). Consistent Nonparametric Regression. *Annals of Statistics, 5*, 595-645.

Watson, G. (1964). Smooth Regression Analysis. *Sankhya Series A, 26*, 359-372.

International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Reviewer Acknowledgements

International Journal of Statistics and Probability wishes to acknowledge the following individuals for their assistance with peer review of manuscripts for this issue. Their help and contributions in maintaining the quality of the journal is greatly appreciated.

Many authors, regardless of whether *International Journal of Statistics and Probability* publishes their work, appreciate the helpful feedback provided by the reviewers.

Reviewers for Volume 7, Number 4

Afsin Sahin, Gazi University, Turkey

Carla J. Thompson, University of West Florida, USA

Encarnación Alvarez-Verdejo, University of Granada, Spain

Felix Almendra-Arao, UPIITA del Instituto Politécnico Nacional, México

Hui Zhang, St. Jude Children's Research Hospital, USA

Luiz Ricardo Nakamura, University of Sao Paulo, Brazil

Mohieddine Rahmouni, University of Tunis, Tunisia

Philip Westgate, University of Kentucky, USA

Sajid Ali, Quaid-i-Azam University, Pakistan

Sohair F. Higazi, University of Tanta, Egypt

Vilda Purutcuoglu, Middle East Technical University (METU), Turkey

Vyacheslav Abramov, Swinburne University of Technology, Australia

Wei Zhang, The George Washington University, USA

Wojciech Gamrot, University of Economics, Poland

Wendy Smith

On behalf of,

The Editorial Board of *International Journal of Statistics and Probability*

Canadian Center of Science and Education

ACADEMIC AND OPEN ACCESS

Canadian Center of Science and Education

CCSENET.ORG

➢ CALL FOR MANUSCRIPTS

International Journal of Statistics and Probability is a peer-reviewed journal, published by Canadian Center of Science and Education. The journal publishes research papers in all areas of statistics and probability. The journal is available in electronic form in conjunction with its print edition. All articles and issues are available for free download online.

We are seeking submissions for forthcoming issues. All manuscripts should be written in English. Manuscripts from 3000–8000 words in length are preferred. All manuscripts should be prepared in LaTeX or MS-Word format, and submitted online, or sent to: ijsp@ccsenet.org

Paper Selection and Publishing Process

a) Submission acknowledgement. If you submit manuscript online, you will receive a submission acknowledgement letter sent by the online system automatically. For email submission, the editor or editorial assistant sends an e-mail of confirmation to the submission's author within one to three working days. If you fail to receive this confirmation, please check your bulk email box or contact the editorial assistant.

b) Basic review. The editor or editorial assistant determines whether the manuscript fits the journal's focus and scope. And then check the similarity rate (CrossCheck, powered by iThenticate). Any manuscripts out of the journal's scope or containing plagiarism, including self-plagiarism are rejected.

c) Peer Review. We use a double-blind system for peer review; both reviewers' and authors' identities remain anonymous. The submitted manuscript will be reviewed by at least two experts: one editorial staff member as well as one to three external reviewers. The review process may take two to four weeks.

d) Make the decision. The decision to accept or reject an article is based on the suggestions of reviewers. If differences of opinion occur between reviewers, the editor-in-chief will weigh all comments and arrive at a balanced decision based on all comments, or a second round of peer review may be initiated.

e) Notification of the result of review. The result of review will be sent to the corresponding author and forwarded to other authors and reviewers.

f) Pay the publication fee. If the submission is accepted, the authors revise paper and pay the publication fee.

g) E-journal is available. E-journal in PDF is available on the journal's webpage, free of charge for download. If you need the printed journals by post, please order at: http://web.ccsenet.org/store.html

h) Publication notice. The authors and readers will be notified and invited to visit our website for the newly published articles.

More Information

E-mail: ijsp@ccsenet.org

Website: http://ijsp.ccsenet.org

Paper Submission Guide: http://submission.ccsenet.org

Recruitment for Reviewers: http://recruitment.ccsenet.org

➢ JOURNAL STORE

To order back issues, please contact the journal editor and ask about the availability of journals. You may pay by credit card, PayPal, and bank transfer. If you have any questions regarding payment, please do not hesitate to contact the journal editor or editorial assistant.

Price: $40.00 USD/copy Shipping fee: $20.00 USD/copy

ABOUT CCSE

The Canadian Center of Science and Education (CCSE) is a private for-profit organization delivering support and services to educators and researchers in Canada and around the world.

The Canadian Center of Science and Education was established in 2006. In partnership with research institutions, community organizations, enterprises, and foundations, CCSE provides a variety of programs to support and promote education and research development, including educational programs for students, financial support for researchers, international education projects, and scientific publications.

CCSE publishes scholarly journals in a wide range of academic fields, including the social sciences, the humanities, the natural sciences, the biological and medical sciences, education, economics, and management. These journals deliver original, peer-reviewed research from international scholars to a worldwide audience. All our journals are available in electronic form in conjunction with their print editions. All journals are available for free download online.

Mission

To work for future generations

Values

Scientific integrity and excellence

Respect and equity in the workplace

CONTACT US

General

Tel: 1-416-642-2606

Fax: 1-416-642-2608

E-mail: info@ccsenet.org

Website: www.ccsenet.org

Mailing Address

1120 Finch Avenue West

Suite 701-309

Toronto, ON., M3J 3H7

Canada

Visiting Address

9140 Leslie St., Suite 110

Richmond Hill, Ontario, L4B 0A9

Canada

The journal is peer-reviewed
The journal is open-access to the full text
The journal is included in:

Aerospace Database
BASE (Bielefeld Academic Search Engine)
EZB (Elektronische Zeitschriftenbibliothek)
Google Scholar
JournalTOCs
Library and Archives Canada
LOCKSS
MIAR
PKP Open Archives Harvester
SHERPA/RoMEO
Standard Periodical Directory
Ulrich's

International Journal of Statistics and Probability

Bimonthly

Publisher	Canadian Center of Science and Education
Address	1120 Finch Avenue West, Suite 701-309, Toronto, ON., M3J 3H7, Canada
Telephone	1-416-642-2606
Fax	1-416-642-2608
E-mail	ijsp@ccsenet.org
Website	http://ijsp.ccsenet.org

ISSN 1927-7032