# PROCEEDINGS BOOK

ICAS CONFERENCE

INTERNATIONAL CONFERENCE ON ADVANCES IN STATISTICS

## 9th International Conference on Advances in Statistics

# ICAS CONFERENCE

INTERNATIONAL CONFERENCE ON ADVANCES IN STATISTICS

June 20-22  2023

http://www.icasconference.com/

**ICAS'2023**

9th International Conference on Advances in Statistics

Published by the ICAS Secretariat

Editors:
**Prof. Dr. Ismihan BAYRAMOGLU**

ICAS Secretariat
Büyükdere Cad. Ecza sok. Pol Center 4/1 Levent-İstanbul
E-mail: aioc.statistics@gmail.com
http://www.icasconference.com

ISBN:

# INTERNATIONAL SCIENTIFIC COMMITTEE

Prof. Dr. Alexei Stepanov
**Immanuel Kant Baltic Federal University – Russia**

Prof. Dr. Ayman BAKLEEZI
**Qatar University – Qatar**

Prof. Dr. Barry C. ARNOLD
**University of California, Riverside – USA**

Prof. Dr. George YANEV
**The University of Texas  Rio Grande Valley – USA**

Prof. Dr. Hamparsum BOZDOGAN
**The University of Tennessee – USA**

Prof. Dr. Hamzeh TORABI
**Yazd University – IRAN**

Prof. Dr. Ilham AKHUNDOV
**University of Waterloo – Canada**

Prof. Dr. Jorge NAVARRO
**Universidad de Murcia – Spain**

Prof. Dr. Jose Maria SARABIA
**University of Cantabria – Spain**

Prof. Dr. Leda MINKOVA
**University of Sofia  – Bulgaria**

Prof. Dr. Narayanaswamy BALAKRISHNAN
**McMaster University – Canada**

Prof. Dr. Nikolai KOLEV
**University of Sao Paulo – Brasil**

Prof. Dr. Sarjinder SINGH
**Texas A&M University-Kingsville – USA**

Prof. Dr. Stelios PSARAKIS
**Athens University of  Economics & Finance – Greece**

Prof. Valerii  B.  Nevzorov
**St. Petersburg State University – Russia**

# LOCAL SCIENTIFIC COMMITTEE

**Prof. Dr. Aydin ERAR**
Mimar Sinan Fine Arts University – Turkey

**Prof. Dr. Baris ASIKGIL**
Mimar Sinan Fine Arts University – Turkey

**Prof. Dr. Coskun KUS**
Selcuk University – Turkey

**Prof. Dr. I. Esen YILDIRIM**
Marmara University – Turkey

**Prof. Dr. Esra AKDENIZ**
Marmara University – Turkey

**Prof. Dr. Gulay BASARIR**
Mimar Sinan Fine Arts University – Turkey

**Prof. Dr. Ismail KINACI**
Selcuk University – Turkey

**Prof. Dr. Mehmet Fedai KAYA**
Selcuk University – Turkey

**Prof. Dr. Mujgan TEZ**
Marmara University – Turkey

**Prof. Dr. Sahamet BULBUL**
Fenerbahce University – Turkey

**Prof. Dr. Yazgi TUTUNCU**
Izmir University of Economics, Turkey

**Assoc. Prof. Dr. Konul KAVLAK**
Bogazici University – Turkey

# ORGANIZATION COMMITTEE

Prof. Dr. Ismihan BAYRAMOGLU
**Izmir University of Economics – Turkey**
Conference Chair

Prof. Dr. Fatma NOYAN TEKELI
**Yıldız Technical University – Turkey**

Prof. Dr. Gulhayat GOLBASI SIMSEK
**Yıldız Technical University – Turkey**

Assoc. Prof. Dr. Aysegul EREM
**Cyprus International University – TRNC**

Assoc. Prof. Dr. Gulder KEMALBAY
**Yıldız Technical University – Turkey**

Assoc. Prof. Dr. Murat OZKUT
**Izmir University of Economics, Turkey**

Instructor PhD Ozlem BERAK KORKMAZOGLU
**Yıldız Technical University – Turkey**

Res. Asst. Ayse BELER
**Izmir University of Economics, Turkey**

Res. Asst. Gokce OZALTUN
**Izmir University of Economics, Turkey**

***Dear Colleagues,***

Dear Colleagues,

On behalf of the Organizing Committee, I am pleased to invite you to participate in 9th International Conference on Advances in Statistics which will be organised online on dates between 20-22 June 2023.

All informations are available in conference web site. For more information please do not hesitate to contact us. info@icascconference.com

We cordially invite prospective authors to submit their original papers to ICAS-2023,

. Applied Statistics
· Bayesian Statistics
· Big Data Analytics
· Bioinformatics
· Biostatistics
· Computational Statistics
· Data Analysis and Modeling
· Data Envelopment Analysis
· Data Management and Decision Support Systems
· Data Mining
· Energy and Statistics
· Entrepreneurship
· Mathematical Statistics
· Multivariate Statistics
· Neural Networks and Statistics
· Non-parametric Statistics
. Operations Research
· Optimization Methods in Statistics
· Order Statistics
· Panel Data Modelling and Analysis
· Performance Analysis in Administrative Process
· Philosophy of Statistics
· Public Opinion and Market Research
· Reliability Theory
· Sampling Theory
· Simulation Techniques
· Spatial Analysis
· Statistical Software
· Statistical Training
· Statistics Education
· Statistics in Social Sciences
· Stochastic Processes
· Supply Chain
· Survey Research Methodology
· Survival Analysis
· Time Series
· Water and Statistics
· Other Statistical Methods

Selected papers will be published in Journal of the Turkish Statistical Association.
https://dergipark.org.tr/en/pub/ijtsa

We hope that the conference will provide opportunities for participants to exchange and discuss new ideas and establish research relations for future scientific collaborations.

Conference Website : https://icasconference.com
E Mail: icasconference.academic@gmail.com

On behalf of Organizing Committee:
Conference Chair
Prof. Dr. İsmihan BAYRAMOĞLU
Izmir University of Economics

# 20 JUNE 2023 TUESDAY

**11:45 – 12:00** **OPENING CEREMONY**
**Professor Ismihan BAYRAMOGLU** / Izmir University of Economics - Turkey

| | |
|---|---|
| **12:00 – 12:30** | **Keynote Speech:**<br>**Professor José Fernando LOPEZ-BLAZQUEZ /** University of Sevilla – Spain<br>*Discrete q-Beta Distributions and Analogs of Order Statistics in Geometric Distributions* |

| | |
|---|---|
| **12:30- 12:45** | **B R E AK** |

## SESSION A

| **SESSION CHAIR** | **Krzysztof JASIŃSKI** | |
|---|---|---|
| **TIME** | **PAPER TITLE** | **PRESENTER** / CO-AUTHOR |
| **12:45 - 13:00** | Optimal Plan for Ordered Step-stress Stage Life Testing | **Shuvashree MONDAL** / Debashis SAMANTA / Debasis KUNDU |
| **13:00 - 13:15** | Sequences of random variables predicted with conditional expectations | **Ismihan BAYRAMOGLU** |
| **13:15 - 13:30** | On the status of component failures in a working coherent system under inspections | **Krzysztof JASIŃSKI** |

| | |
|---|---|
| **13:30 - 13:35** | **B R E AK** |

| | |
|---|---|
| **13:35 - 13:55** | **Invited Speech:**<br>**Professor Anna DEMBIŃSKA** / Warsaw University of Technology – Poland<br>*Likelihood Inference for Geometric Lifetimes of Components of k-out-of-n Systems* |

| | |
|---|---|
| **13:55 - 14:00** | **B R E AK** |

## SESSION B

| **SESSION CHAIR** | **Ozge KURAN** | |
|---|---|---|
| **TIME** | **PAPER TITLE** | **PRESENTER** / CO-AUTHOR |
| **14:00- 14:15** | Efficient Sampling from the PKBD Distribution | **Lukas SABLICA** / Kurt HORNIK / Josef LEYDOLD |
| **14:15 - 14:30** | Robust Estimation of Correlation Coefficient via Ranked Set Sampling: Application to Bivariate Normal Distribution | **Yusuf Can SEVIL** / Tugba OZKAL YILDIZ |
| **14:30 - 14:45** | A New Two-Parameter Biased Prediction Class in Linear Mixed Models | **Ozge KURAN** |

| 14:45 – 14:50 | BREAK |
|---|---|

| 14:50 - 15:20 | **Keynote Speech:**<br>**Professor Leda MINKOVA** / Sofia University – Bulgaria<br>*Compound Discrete Time Geometric Process* |
|---|---|

| 15:20 - 15:30 | BREAK |
|---|---|

## SESSION C

| SESSION CHAIR | Jamila KALANTAROVA | |
|---|---|---|
| **TIME** | **PAPER TITLE** | **PRESENTER** / CO-AUTHOR |
| 15:30 - 15:45 | Benford's Law and the Digits of Powers of Two in Ternary Numeral System | **Yagub N. ALIYEV** |
| 15:45 - 16:00 | Gegenbauer Wavelet Solutions of Lane-Emden Equations | **Demet OZDEK /** Sevin GUMGUM |
| 16:00 - 16:15 | Global non-existence of the initial value problem for nonlinear thermoelasticity-type equations | **Jamila KALANTAROVA** |

| 16:15 - 16:20 | BREAK |
|---|---|

| 16:20 – 16:40 | **Invited Speech:**<br>**Professor Maria LONGOBARDI** / Università di Napoli Federico II – Italy<br>*Cumulative entropies in terms of moments of order statistics* |
|---|---|

| 16:40 - 16:45 | BREAK |
|---|---|

## SESSION D

| SESSION CHAIR | Gulder KEMALBAY | |
|---|---|---|
| **TIME** | **PAPER TITLE** | **PRESENTER** / CO-AUTHOR |
| 16:45 – 17:00 | Skewness Correction Ewma for Non-Normal Processes | **Aminou MOUSTAPHA** / Derya KARAGÖZ |
| 17:00 – 17:15 | Bayesian Inversion into Soil Types with Kernel-Likelihood Models | **Selamawit SERKA MOJA** / Henning MORE |
| 17:15 – 17:30 | Computations of Conditional Expectations for General Models of Ordered Random Variables | **Ege CAGATAY /** Ismihan BAYRAMOGLU |
| 17:30 - 17:45 | Distribution of conditional bivariate order statistics via modifications of the bivariate binomial distribution | **Gulder KEMALBAY** |

| 17:45 – 18:00 | BREAK |
|---|---|

| 18:00 – 18:30 | **Keynote Speech:**<br>**Professor Narayanaswamy BALAKRISHNAN** / McMaster University – Canada<br>*Relationships between cumulative entropy/extropy, Gini mean difference and probability weighted moments* |
|---|---|

| 18:30 – 18:45 | **BREAK** |
|---|---|

| 18:45 – 19:15 | **Keynote Speech:**<br>**Professor Barry C. ARNOLD** / University of California- USA<br>*Half-normal curiosities* |
|---|---|

# 21 JUNE 2023 WEDNESDAY

| 11:45 – 12:15 | **Keynote Speech:**<br>**Professor Eugenia STOIMENOVA** / Bulgarian Academy of Sciences – Bulgaria<br>*Multi-Sample Rank Tests for Location against Ordered Alternatives* |
|---|---|

| 12:15 - 12:30 | **BREAK** |
|---|---|

## SESSION E

| SESSION CHAIR | Fatma NOYAN TEKELI | |
|---|---|---|
| **TIME** | **PAPER TITLE** | **PRESENTER** / CO-AUTHOR |
| 12:30 – 12:45 | Drought Area Prediction with Remote Sensing Data Based on Machine Learning Methods | **Gokce GOK/** Burcu HUDAVERDI |
| 12:45 – 13:00 | Online Statistics Education Using Web-based Software, eStat | **Jung Jin LEE** |
| 13:00 – 13:15 | Implementation of Recommender Systems Using Genetic Algorithm for E-Commerce | **Merve POSLU** / Guvenc ARSLAN / Sevil BACANLI |
| 13:15 – 13:30 | Daily Natural Gas Price Prediction with Long Short-Term Memory | **Coskun PARIM** / Batuhan OZKAN / Fatma NOYAN TEKELI |

| 13:30- 13:40 | **BREAK** |
|---|---|

## SESSION F

| SESSION CHAIR | Gulhayat GOLBASI SIMSEK | |
|---|---|---|
| **TIME** | **PAPER TITLE** | **PRESENTER** / CO-AUTHOR |
| 13:40 - 13:55 | Analysis of Maximum Precipitation in Thailand Using Model Averaging of Non-Stationary Extreme Value Models | **Thanawan PRAHADCHAI /** Jeong-Soo PARK |

|  |  |  |
|---|---|---|
| **13:55 - 14:10** | Machine Learning Applications in Detecting Obfuscated Malware | **Opetunde IBITOYE** / Mary AGOYI / Aysegul EREM |
| **14:10 – 14:25** | Analyzing Centrality and Connectivity in Airport Networks: A Focus on Turkey and Europe | **Cem ERSOZ** / Filiz KARAMAN |
| **14:25 – 14:40** | A Hybrid Method Based on Wavelet Transform and Neural Network Approach for Meteorological Data Prediction | **Gokce Nur TASAGIL ARSLAN /** Serpil KILIC DEPREN |
| **14:40 - 14:55** | Stochastic Processes in Insurance Science | **Andreas MAKRIDES** |

|  |  |
|---|---|
| **14:55 - 15:15** | **BREAK** |

|  |  |
|---|---|
| **15:15–15:45** | **Keynote Speech:**<br>**Professor Haikady NAGARAJA/** The Ohio State University – USA<br>*Odds ratios from logistic, geometric, Poisson, and negative binomial regression models* |

|  |  |
|---|---|
| **15:45 - 16:00** | **BREAK** |

## SESSION G

| **SESSION CHAIR** | **Arzu BAYGUL EDEN** |  |
|---|---|---|
| **TIME** | **PAPER TITLE** | **PRESENTER** / CO-AUTHOR |
| **16:00 – 16:15** | Unleashing Diversity Potential: The What-If Group Voting Recommendation System | **Uguray DURDU** / Atilla Recep BASARAN / Alperen BALIK / Enes KACAR / Alper ONER |
| **16:15 - 16:30** | A New Perspective on the Evaluation of Comorbidity Indices on Survival | **Alev BAKIR** / Benan MUSELLIM / Mustafa S. SENOCAK |
| **16:30 – 16:45** | Turkish Validation of Meta-analysis Observational Studies in Epidemiology (MOOSE) Check List | **Arzu BAYGUL EDEN** / Neslihan GÖKMEN INAN / Alev BAKIR KAYI |

|  |  |
|---|---|
| **16:45 - 17:00** | **BREAK** |

|  |  |
|---|---|
| **17:00 - 17:30** | **Keynote Speech:**<br>**Professor Omer OZTURK** / The Ohio State University – USA<br>*Order Restricted Randomized Block Designs* |

|  |  |
|---|---|
| **17:30 - 18:00** | **Keynote Speech:**<br>**Professor Ilham AKHUNDOV** / University of Waterloo – Canada<br>*Characterizations of distributions based on regression and related statistics* |

|  |  |
|---|---|
| **18:00 - 18:15** | **BREAK** |

## SESSION H

| SESSION CHAIR | Yazgı TUTUNCU | |
|---|---|---|
| **TIME** | **PAPER TITLE** | **PRESENTER** / CO-AUTHOR |
| **18:15 – 18:30** | Statistical Analysis of Altitude Effect on Daily Selective Attention Task Performance | **Yazgı TUTUNCU /** Sermin TUKEL / Hasan KAZDAGLI |
| **18:30 – 18:45** | A Numerical Approcah to Investigate the Spread of COVID-19 | **Ayse BELER /** Sevin GUMGUM |

# 22 JUNE 2023 THURSDAY

| | |
|---|---|
| **09:00 – 09:30** | **Keynote Speech:**<br>**Professor Alex KARAGRIGORIOU** / University of the Aegean – Greece<br>*Power Divergence Statistics and the Modified Family for the Zero Count Case* |

| | |
|---|---|
| **09:30 - 09:40** | **BREAK** |

| | |
|---|---|
| **09:40 - 10:00** | **Invited Speech:**<br>**Professor Coskun KUS** / Selcuk University – Turkey<br>*A Class of Non-parametric Tests for the Two-Sample Problem Based on Order Statistics* |

| | |
|---|---|
| **10:00 - 10:10** | **BREAK** |

| | |
|---|---|
| **10:10 - 10:30** | **Invited Speech:**<br>**Professor Debasis KUNDU** / Indian Institute of Technology Kanpur – India<br>*Bivariate Distributions with Singular Components* |

**10:30 – 10:45 CLOSING CEREMONY**

**Professor Ismihan BAYRAMOGLU** / Izmir University of Economics - Turkey

# Digits of Powers of 2 in Odd Based Numeral System and Benford's Law

*Yagub N. ALIYEV* [1],

[1]School of IT and Engineering, ADA University

Ahmadbey Aghaoglu str. 61, Baku, AZ1008, Azerbaijan

e-mail: yaliyev@ada.edu.az

**Abstract**

We study the digits of the powers of 2 in p=2k+1 based number system. We propose an easy algorithm for doubling numbers in this number system. Using this algorithm, we explain the appearance of "stairs" of 2k's and 0's when the number 2^(n+1) is written on top of 2^n (n=0,1,2,…) in a natural way so that for example the last digits are forming one column, the pre-last digits are forming another column, etc. We also look at the patterns formed by the first digits, the patterns formed by the last digits and use this to prove that the sizes of these "stairs" as blocks of 0's and 2k's are unbounded. We also discuss how this regularity changes when the digits move from left end of the numbers to the right end.

***Key Words:*** *Ternary, digits, zeros, Benford's law, powers of two.*

## 1. Introduction

Let us write first powers of two $(1,2,4, \ldots, 2^{53})$ in, for example, quinary numeral system (k=2) so that their first digits are aligned along one vertical column. Note that the digits with higher place values are written after the digits with lower place values. For example, $211 = 2 + 1 \cdot 5 + 1 \cdot 5^2$.

1

2

**4**

31

13

211

**4**22

3001

1102

220**4**

**44**031

3**4**113

1**4**3211

232032

**4**101101

3302202

121**44**0**4**

2**4**23**4**131

**4**302**4**313

321**4**32311

103320232

201201**4**101

**4**02**4**023302

31**4**31**4**121**4**

1332333**4**231

2120222**4**023

**4**240**444**31411

30413**44**23332

11331**44**022201

221233**4**144402

**44**2**4**12**4**33**44**1**4**

3**4**043**4**322**44**331

1**4**132**4**200**44**223

23310**4**0103**4**0021

**4**12303120**4**...

| | | |
|---|---|---|
| 2130233223**4**003**4**201**4**1 | 111022**4**103221**4**13**4**33**4**2 | 3**44**11**4**003**4**002**44**121033**4** |
| **4**211**4**12002**4**101**4**01233 | 2220**44**3301003331**4**22**4**01 | 1**44**323101**4**10**4**3**4**3**4**2012**4**1 |
| 303233**4**00**4**3302312**4**121 | **444**03**4**2212001223300**4**12 | 23**4**20230233032**4**2**4**012**4**33 |

We can easily observe several interesting patterns from this table.

Observation I. If we look only at the first $k$ digits of each power of two, then, as we go downwards along the table, we can see that all possible $k$-digit blocks (some are impossible) appear in the table and they repeat periodically. For example, if $k$=2, then there are $4 \cdot 5^{k-1} = 20$ possible 2-digit blocks and all of them appear in the first 3 columns of the above table and would appear periodically if we would continue to extend the table upwards.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 40 | 13 | 42 | 11 | 44 | 14 | 41 | 12 | 43 |
| 20 | 31 | 21 | 30 | 22 | 34 | 23 | 33 | 24 | 32 |

Observation II. If we look only at the last n digits, then we can see that all the possible n-digit blocks (excluding those which start with 0) appear but this time not periodically. For example, if n=2 then there are 20 possible $2$-digit blocks and almost all of them appear in the table, although not with the same frequency. One can notice that the numbers 01 or 11, for example, appear more frequently than, say, 14 or 24. The number 44 doesn't appear at all, but it appears in 2^58. Similarly, 03 doesn't appear at first but it appears at 2^55 and 2^62.

Observation III. There are blocks of 0's and 4's in the shape of stairs, larger ones of which seem to have sizes which increase unboundedly. This means that there are arbitrarily large such stairs. Each stair of the stairs is of height either 1, 2, or 3 digits.

We will see that both observations I and II are in general true and both can be used to explain III.

## 2 The doubling algorithm

Let us first describe the algorithm for doubling an arbitrary positive integer in p=2k+1 based number system. Consider the substitutions

$$A = \begin{pmatrix} 0 & 2 & 4 & \ldots & 2k & 1 & 3 & 5 & \ldots & 2k-1 \\ 0 & 1 & 2 & \ldots & k & k+1 & k+2 & k+3 & \ldots & 2k \end{pmatrix} B = \begin{pmatrix} 0 & 2 & 4 & \ldots & 2k & 1 & 3 & 5 & \ldots & 2k-1 \\ k & k+1 & k+2 & \ldots & 2k & 0 & 1 & 2 & \ldots & k-1 \end{pmatrix}.$$

1) Add an extra 0 to the left of the given number. Start with the first digit of the number on the left end of it and apply $A$. Write the result immediately below this digit.

2) If we obtained 0, 2 or 4, then for the next digit of the number to the right, we use the last used substitution, otherwise if we obtained 1 or 3 then we switch to the other substitution, apply it for the next digit on the right. Write the result immediately below the digit.
3) Return to step 2) unless you already reached the extra 0 at the right end.

After completion of the algorithm, we don't write that extra zero at the left of the resulting numbers and shift the numbers so that the leftmost digits are on the same column. The proof of this algorithm is simple, and we skip it. Let us see how this algorithm is applied to a string of fours and zeros in a number when k=2. Suppose that we have a string of zeros as in the red part of the numbers in the following list.

$$\begin{array}{lcccccc}
2^{37} = \cdots & 4 & 2 & 2 & 2 & 4 & \ldots \\
2^{38} = \cdots & 4 & \boxed{0} & \boxed{0} & \boxed{0} & 4 & \ldots \\
2^{39} = \cdots & 3 & 1 & \boxed{0} & \boxed{0} & 3 & \ldots \\
2^{40} = \cdots & 2 & 3 & \boxed{0} & \boxed{0} & 1 & \ldots \\
2^{41} = \cdots & 0 & 2 & 1 & \boxed{0} & 2 & \ldots \\
2^{42} = \cdots & 1 & 4 & 2 & \boxed{0} & 4 & \ldots \\
2^{43} = \cdots & 3 & 3 & 0 & 1 & 3 & \ldots
\end{array}$$

According to the described doubling algorithm, the digit 0 can be obtained only from 0 (the substitution $A$) or from 2 (the substitution $B$). In both cases, we keep using the same substitution until we run out of zeros. Because of this, either

1. all these zeros are obtained from only twos (e.g. check above how the red digits of 2^38 are obtained from 2^37),
2. or all these zeros are obtained from again the zeros (e.g. check above how the red digits of 2^39 are obtained from 2^38.

These observations that concerned appearance of zeros in for base 5 numeral system, can be extended for fours in base 5 but also generalized for appearance of zeros and (2k)s for any other bases p=2k+1.

## 3 The first digits

We will continue with the explanation of I Observation. We need some lemmas to continue.

**Lemma 1.** If $k$ is any positive integer then $5^k | \left(2^{2\cdot 5^{k-1}} + 1\right)$, but $5^{k+1} \nmid \left(2^{2\cdot 5^{k-1}} + 1\right)$.

**Proof.** This can be proved using the method of mathematical induction. Denote $A_k = 2^{2\cdot 5^k} + 1$. For $k = 1$ we have $5^1 | A_0$, but $5^2 \nmid A_0$. Suppose that it is true for $k = n$, that is $5^{n+1} | A_n$, but $5^{n+2} \nmid A_n$. Then $A_{n+1} = 2^{2\cdot 5^{n+1}} + 1 = \left(2^{2\cdot 5^n}\right)^5 + 1 = \left(2^{2\cdot 5^n} + 1\right)\left(\left(2^{2\cdot 5^n}\right)^4 - \left(2^{2\cdot 5^n}\right)^3 + \left(2^{2\cdot 5^n}\right)^2 - 2^{2\cdot 5^n} + 1\right) =$
$A_n((A_n - 1)^4 - (A_n - 1)^3 + (A_n - 1)^2 - A_n + 2) = A_n(A_n^4 - 5A_n^3 + 10A_n^2 - 5A_n + 5)$. Note that $5 | (A_n^4 - 5A_n^3 + 10A_n^2 - 5A_n + 5)$ but $5^2 \nmid (A_n^4 - 5A_n^3 + 10A_n^2 - 5A_n + 5)$. Therefore $5^{n+2} | A_{n+1}$, but $5^{n+3} \nmid A_{n+1}$.

**Lemma 2.** If $k$ is any positive integer then $5^k|2^{4\cdot 5^{k-1}}-1$, but $5^{k+1}\nmid 2^{4\cdot 3^{k-1}}-1$.

**Proof.** Since $2^{4\cdot 5^{k-1}}-1=\left(2^{2\cdot 5^{k-1}}-1\right)\left(2^{2\cdot 5^{k-1}}+1\right)$ and $5\nmid\left(2^{2\cdot 5^{k-1}}-1\right)$, this easily follows from Lemma 1.

**Note.** We could also use the special case of Euler's Theorem, which says that $2^{\varphi(5^{n+1})}\equiv 1\ (mod\ 5^{n+1})$, and the fact that $\varphi(5^{n+1})=4\cdot 5^n$. Here $\varphi$ is The Euler Phi-Function [6, sec. 6.3 and 7.1]. Also note that $\operatorname{ord}_{5^{n+1}}2=4\cdot 5^n$, which means that $x=4\cdot 5^n$ is the least positive integer such that $2^x\equiv 1(\operatorname{mod} 5^{n+1})$. Otherwise, since $\operatorname{ord}_{5^{n+1}}2|\varphi(5^{n+1})$ and $\varphi(5^{n+1})=4\cdot 5^n$, either (I option) $\operatorname{ord}_{5^{n+1}}2=5^k$ for some $0\le k\le n$, (II option) $\operatorname{ord}_{5^{n+1}}2=2\cdot 5^k$ for some $0\le k\le n$, or (III option) $\operatorname{ord}_{5^{n+1}}2=4\cdot 5^k$ for some $0\le k<n$. But $5\nmid\left(2^{5^k}-1\right)$ and $5\nmid\left(2^{2\cdot 5^k}-1\right)$, as mentioned earlier, so, the first and second options are not possible. The third option is also impossible, because in this case $2^{4\cdot 5^k}\equiv 1(\operatorname{mod} 5^{n+1})$ for some $0\le k<n$. But as proved in Lemma 2 above $5^{k+1}\nmid 2^{4\cdot 5^k}-1$, therefore $5^{n+1}\nmid 2^{4\cdot 5^k}-1$, too. The equality $\operatorname{ord}_{5^{n+1}}2=\varphi(5^{n+1})$ that we just proved means that 2 is a primitive root modulo $5^{n+1}$. See [6, sec. 9.3, p. 327]. So, we proved the following theorem that explains I Observation above.

**Theorem 1.** Except those which start by zero, any finite sequence of digits can appear infinitely many times, at the end of a quinary number system notation of a power of 2.

**Note.** If $n$ is a positive integer, then $\{1,2,2^2,2^3,\dots,2^{\varphi(5^{n+1})-1}\}$ gives the set of $\varphi(5^{n+1})$ integers such that each element of the set is relatively prime to 5, and no two different elements of the set are congruent modulo $5^{n+1}$, i.e. the set forms a reduced residue set modulo $5^{n+1}$. So, the last theorem says that if the elements of the set $\{1,2,2^2,2^3,\dots,2^{\varphi(5^{n+1})-1}\}$ are written in base 5 number system, then the first $n$ digits go through all possible $n$-tuples without repetitions, except those with 0 at the start. This means that if we will look only at the first $k$ digits, then with sufficiently large $n$, we can get any sequence of the digits 0, 1, 2, 3, and 4, equally frequently, including arbitrary large number of consecutive zeros $(00\dots 0)$ or fours $(44\dots 4)$, provided that $k$ is also taken sufficiently large. Because of the doubling algorithm described above, any such block of zeros (or fours) is included in a triangular like structure of zeros (or twos). This is a convincing argument for the proof of III Observation which claimed that the dimensions of such structures are unbounded. But this doesn't say anything about 1) how frequent such structures appear or 2) how fast they grow in size. The last two questions are probably deeper questions and require more thorough analysis.

This theorem can be generalized for arbitrary bases $p=2k+1$. But for some bases some of the starting digits are not possible. For example, if $p=7$ then the digits 3, 5 and 6 are not possible and if $p=9$ then the digits 3 and 6 are not possible. But still the frequency of possible blocks of digits is constant.

## 4 The last digits

Let us now turn our attention to the first digits of the elements of the set $\{1, 2, 2^2, 2^3, \dots\}$, when they are written in base $p = 2k + 1$ number system. Suppose that the first $m$ digits of $2^n$ are $(\overline{a_1 a_2 \dots a_m})_p = A$, where $a_1 \in \{1,2,\dots,2k\}$ and $a_i \in \{0,1,2,\dots,2k\}$ for $i = 2, \dots, m$. Then

$$A \cdot p^l < 2^n < (A+1) \cdot p^l,$$

for some nonnegative integer $l$. Taking base $p$ logarithm of both sides gives

$$l + \log_p A < n \log_p 2 < l + \log_p (A+1).$$

Since $m - 1 \le \log_p A < m$ and $m - 1 < \log_p (A+1) \le m$, we conclude that

$$l + m - 1 < n \log_p 2 < l + m.$$

This means that $l + m - 1$ is simply the integer part of $n \log_p 2$. So,

$$\log_p A - m + 1 < n \log_p 2 - \lfloor n \log_p 2 \rfloor < \log_p (A+1) - m + 1.$$

Note that $\left[\log_p A - m + 1, \log_p (A+1) - m + 1\right] \subseteq [0,1]$. By the well-known result of Bohl [2], Sierpinski [4], and Weyl [3] (see also [5, Chapter 1, Example 2.1]) the sequence $\{n \log_p 2\}(n = 1,2,\dots)$ is uniformly distributed modulo 1. In particular, this means that there are infinitely many $n$ such that the difference $n \log_p 2 - \lfloor n \log_p 2 \rfloor$ is in the interval $\left[\log_p A - m + 1, \log_p (A+1) - m + 1\right]$. This proves the following theorem that explains II Observation above.

**Theorem 2.** Except those which end by zero, any finite sequence of digits can appear infinitely many times, at the end of $p = 2k + 1$ based number system notation of a power of 2.

**Note.** Since arbitrary large number of consecutive zeros $(00 \dots 0)$ or twos $\left((2k)(2k) \dots (2k)\right)$ should also appear infinitely many times we come back again to Observation III, with a completely new explanation for it. Suppose that $m$ is a positive integer and $n = n_0$ is the least integer such that $2^n > p^m$. Then the first $m$ digits of $\{2^n\}(n = n_0, n_0 + 1, \dots)$, give all of possible $(p-1) \cdot p^{m-1}$ sequences of digits at the end of these numbers. The frequency with which $\overline{a_1 a_2 \dots a_m} = A$ appear at the end when $n \to \infty$, is equal to the length of the interval $\left[\log_p A - m + 1, \log_p (A+1) - m + 1\right]$, which is $\log_p \frac{A+1}{A} = \log_p (A+1) - \log_p A$. In contrast to the case of last digits described above, where the frequency is the same for all combinations, the frequency of the last digits show preference for smaller $A$, when $m$ is fixed. This phenomenon is generally known as Benford's law. See e.g. Exercise 20.3.2 in [9, p. 502] for a version of this law similar to ours.

**Theorem 3.** If the powers of 2 are written so that each next power of 2, in ternary number system notation, is written below of the previous power of 2, and the unit digits are all on the same vertical line, then arbitrarily large triangular blocks of zeros and (2k)s can appear in this infinite triangular table.

In view of these results, it would be interesting to study the question of frequency for the interior digits of the powers of two and how this frequency changes when the block of digits $A$ shift from left endpoint, where the frequencies are different and obey Benford’s law, to the right endpoint where all the frequencies are equal.

There is also a related but specific question by P. Erdős: how frequently do the powers of 2 have ternary expansions that omit the digit 2? He conjectured that this holds only for finitely many powers of 2. See [1] for the discussion of this problem. In view of the results of the current paper, Erdős's conjecture can be generalized in the following way. There are only finitely many powers of 2 which does not intersect the triangular structures containing only $(2k)$s in 2k+1 based numeral system.

A similar observation can be made about the powers of two which miss the regions of zeros in any odd based numeral system. Better understanding of these structures of zeros and twos can be useful for the future solution of Erdős's problem. Ternary numeral system version of the question was discussed in [10].

Base 7

1
2
4
11
22
44
121
242
415
1331
2662
4565
14641
21623
42556
153461
230263
4604501
1611412
2532134
43052611

10135532
20263405
40450213
111414262
222131555
4442623441
1225556123
2443446346
4120226026 1
13404451453
26111243140 1
45322410311 2
14054130622 4
21132360554 11
4226465134232
1545264361564
23245520633621
46414450500652
16231241310545
25562413620324 1
43465136550641 3

102643654415236
204520642233565 1
401450525466364 3
1121413532660620 1
2242136305561550 2
4415265013463341 4
1233554126160023 11
2466342345351046 22
4166025614043016 54
1356143631110125421
2636310603220243252
4506030516440410535
14151603352211303041
21333516635422606013
42666335604254515126
155666004611532433345 1
23466101632305100024 3
461663025056033000410 1
163560143136166000130 2
250461210365356100260 4
43116342065404630004511 1

| | | |
|---|---|---|
| 10325025054211601014322 | 113506206520420021150642 | 1241661364134620232656303 |
| 20643143132522512021054 | 226315505450150042231525 | 2413563652361650464546016 |
| 405203103643543340420321 | 4450334132412310154623531 | |

Base 9

| | | |
|---|---|---|
| 1 | 50050761 | 720557220531603 |
| 2 | 11011543 | 550126540172316 |
| 4 | 22022107 | 1212432012556231 |
| 8 | 440442051 | 2424864024124562 |
| 71 | 880884013 | 4848740148248145 |
| 53 | 781780126 | 87086012875873811 |
| 17 | 5835812431 | 76174124762867732 |
| 251 | 1871834862 | 54350348545746674 |
| 413 | 2763778745 | 107116872026044601 |
| 826 | 45475686021 | 205323865043188312 |
| 7531 | 81062484142 | 401746742186278724 |
| 5272 | 73035870384 | 802504605274558658 |
| 1555 | 570618616701 | 7141183115502284281 |
| 21221 | 161337433612 | 5382277221214470573 |
| 42442 | 233665076334 | 1775456542428851167 |
| 84884 | 466342154768 | 25620242058478232361 |
| 708801 | 8347842106481 | 41450484018068564633 |
| 517812 | 7785805203083 | 82811870127138240476 |
| 135834 | 5682811506077 | 757327612453675808541 |
| 261878 | 14857321131561 | 526745434817362817203 |
| 4337681 | 28726742262143 | 153602078735735735506 |
| 8665483 | 47653605435286 | 2173140586716716712131 |
| 7442087 | 854273110715741 | 4257280184633633634262 |

8416571270476376378435 7033263451854754758 0711

## Acknowledgements

This work was completed with the support of ADA University Faculty Research and Development Fund. Many thanks to ADA University and especially School of IT and Engineering for their constant support. This paper has been presented in ICAS 2023 Online on dates between 20 – 22 June 2023.

## References

[1] Lagarias, J.C. (2009), Ternary expansions of powers of 2. Journal of the London Mathematical Society, 79: 562-588. https://doi.org/10.1112/jlms/jdn080

[2] P. Bohl, (1909) Über ein in der Theorie der säkutaren Störungen vorkommendes Problem, J. reine angew. Math. 135, pp. 189–283.

[3] Weyl, H. (1910). "Über die Gibbs'sche Erscheinung und verwandte Konvergenzphänomene". Rendiconti del Circolo Matematico di Palermo. 330: 377–407. https://doi.org/10.1007/BF03014883

[4] W. Sierpinski, (1910) Sur la valeur asymptotique d'une certaine somme, Bull Intl. Acad. Polonaise des Sci. et des Lettres (Cracovie) series A, pp. 9–11.

[5] Kuipers, L. and Niederreiter, H. (1974) Uniform Distribution of Sequences. John Wiley; Russian translation: Nauka, 1985.

[6] K.H. Rosen, Elementary Number Theory and its Applicatons, 4th edition, Addison-Wesley-Longman, 2000.

[7] P. Strzelecki, On powers of 2, EMS Newsletter, 52 (2004) 7-8; translated from Polish journal Delta 7 (1994).

[8] I.M. Sobol, Points uniformly filling a high-dimensional cube, Znaniye, Mathematics, Cybernetics series, 2 (1985), 1-32. (in Russian)

[9] A. Klenke, (2020) Probability Theory, A Comprehensive Course, Universitext, 716 p. https://doi.org/10.1007/978-3-030-56402-5

[10] Y. N. Aliyev, (2023). Digits of powers of 2 in ternary numeral system. Notes on Number Theory and Discrete Mathematics, 29(3), 474-485, DOI: 10.7546/nntdm.2023.29.3.474-485

# Online Statistics Education Using Web-based Software, eStat

*Jung Jin LEE*

Professor, ADA University, Baku, Azerbaijan, jlee@ada.edu.az

**Abstract**

Well known statistical packages such as SAS, SPSS and R are widely used for data processing, but they have not paid attention much to develop modules specialized in Statistics Education. eStat is a free web-based graphical software at www.estat.me which has been developed especially for online Statistics Education. eStat is designed to practice all topics of Introductory Statistics at undergraduate level with many simulation modules, but also to practice data processing online. More specifically, this software includes the modules of all statistical distributions, Law of Large Number, Central Limit Theorem, Confidence Interval, Testing Hypothesis, Analysis of Variance, Nonparametric Tests, Regression Analysis and Forecasting. Moreover, eStat is linked with an eBook which includes many examples, presentation files and lecture movies for students to comprehend the subject of Statistics. This integrated system of eStat was adopted for online teaching by ADA University during pandemic period and it has been very useful for both students and lecturers. It is now used for offline teaching. This paper introduces useful experience of eStat for Statistics Education.

***Key Words:*** **Statistics education; statistical package; web-based software; eStat**

## 2. Introduction

Recent advance in computer technology enabled to develop statistical packages such as SAS, SPSS, R for massive data processing and these packages are widely used in data analysis. However, these well-known packages have not paid their attention to develop modules suitable for teaching Statistics. Many individual developers have developed software for Statistics Education, but most of them are limited to some specific topics or focused on a certain level of students. Current information society has made Statistics Education popular at elementary, middle, high school, and many majors at university. Hence, it is highly vital to develop a software which can be used for teaching Statistics to all levels of students. The Royal Statistical Society initiated Statistics Education for school children by using the census@school project [1]. Clifford Konold and Craig Miller at the University of Massachusetts Amherst developed a software called TinkerPlots [5] which enables exploratory data analysis and modeling for the use of students in grades 4 through university. TinkerPlots is a good tool for training logical ways of thinking overall, but it does not include many statistical analyses.

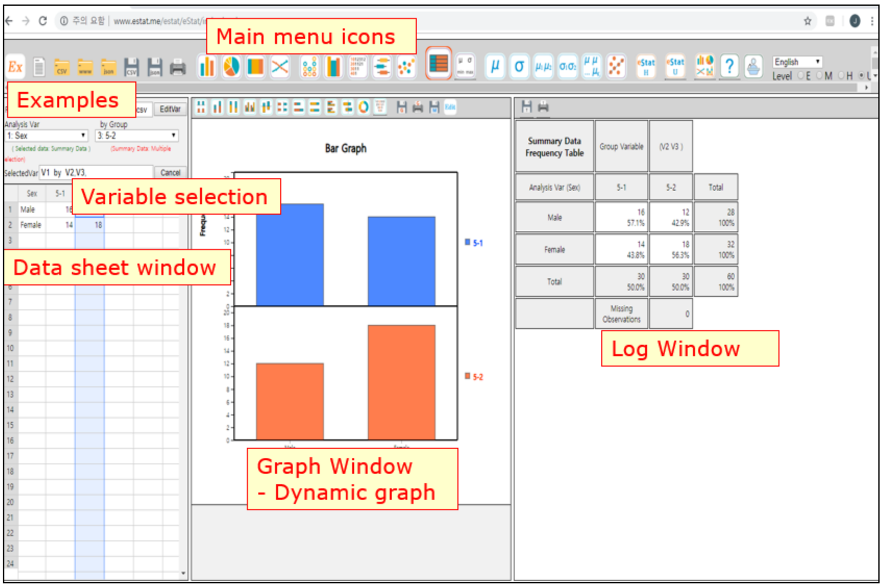
Considering this shortcoming in the education of Statistics, Jung Jin Lee and seven other professors [4] developed a PC based software, **Tong-Gramy, in 2015 for the use of students in elementary and middle school with** partial support of Korean government. Following this, Jung Jin Lee and seven other professors developed a web-based software called eStat in 2019 [2] which was written in Korean, and it has been upgraded until now to the international version translated by 20 languages. eStat at www.estat.me is a free web-based software that includes all statistical analysis for the use of students at middle school, high school and university. eStat includes not only many dynamic graphs for visualizing data with an easy user interface, but also many simulation and data processing modules for statistical analysis with appropriate graphs. Furthermore, eStat includes modules of all statistical distributions such as Binomial, Normal, t, chi-square etc., Law of Large Number, Central Limit Theorem, Confidence Interval, Testing Hypothesis, ANOVA, Nonparametric Tests, Regression and Forecasting.  eStat also includes an eBook that is linked with presentation files, lecture movies, practicing modules and each example of eBook is linked directly with a practicing module. eStat has now become an integrated system of Statistics Education for use in anytime and anywhere if the internet is available.

eStat has been used for online teaching at ADA University in Azerbaijan for two semesters of Business Statistics courses during pandemic period and it has been useful for both students and lecturers in terms of the quality of educational experience. eStat is now used for offline teaching of Statistics. This paper shares useful experience of eStat for Statistics Education at university level. The overall design structure of eStat is introduced in section 2. Useful features of eStat are explained in section 3. Section 4 concludes this paper.

### 3. eStat System

eStat is a web-based, dynamic graphical software with an easy user interface. eStat includes data processing by web and many statistical analyses with easy graphical outputs for understanding. eStat has been developed to work in any operating system of PC, Tablet, or mobile device with any web browser and therefore users can utilize eStat anytime and anywhere if the internet is available. eStat has been developed by using HTML5, CSS3, JavaScript [3] and it works 100% on a browser which follows the standard grammar of HTML5 such as Google Chrome. eStat is also working on the most of other browsers which do not follow the standard grammar of HTML5 such as Microsoft Edge, but certain function such as saving a file might not work.

Figure 1 is the main screen of eStat which can be viewed by typing the web address, www.estat.me, at a browser of your device.



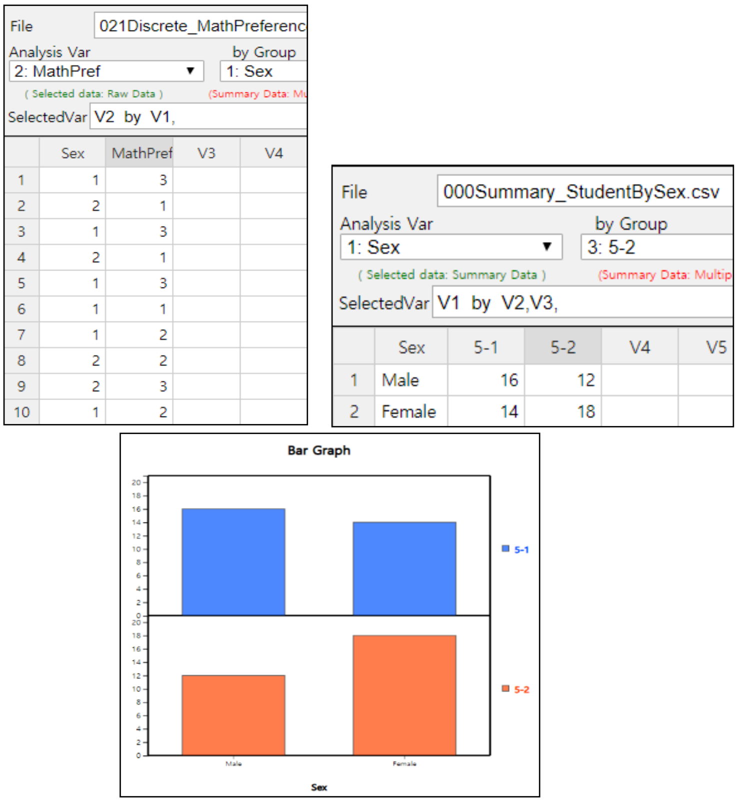
| Summary Data Frequency Table | Group Variable | (V2 V3 ) | |
|---|---|---|---|
| Analysis Var (Sex) | 5-1 | 5-2 | Total |
| Male | 16<br>57.1% | 12<br>42.9% | 28<br>100% |
| Female | 14<br>43.8% | 18<br>56.3% | 32<br>100% |
| Total | 30<br>50.0% | 30<br>50.0% | 60<br>100% |
| | Missing Observations | 0 | |

Figure 1. Structure of main screen of eStat, www,estat.me

The main screen consists of three windows, Data-sheet Window, Graph Window and Log Window. There is Main menu icons at the top of the screen and Sub-menu icons in each window. This design of menu icons is adopted to help mobile users and/or school children who are not accustomed to a scroll-down menu. If a user enters data at Data-sheet Window as in Figure 1, click Bar Graph icon for analysis, and select variables for analysis as V1, V2, V3 by mouse clicking (or finger pointing in case of mobile device) on each variable name, then the bar graph will appear at Graph Window. If you click the icon of Frequency Table, a frequency table will appear at Log Window. You can change the bar graph with other graph style by clicking any Sub-menu icon located at the above of Graph Window. Many examples can be found by clicking Ex icon located at the upper left-hand side of the main screen. Language and Level of study can be selected from the menu located at the upper right-hand side of the main screen.

Entered data in Data-sheet Window can be saved as a CSV (comma separated value) or JSON file format at your local PC by using icons of CSV Save or JSON Save. Any data in CSV or JSON format can be loaded by using icons of CSV Open, www Open and JSON Open from a local PC or any server in the world. The graph drawn in Graph Window can be saved as a png file format at the download folder of a local PC and the tables created in Log Window can be saved as a html file format at the download folder of a local PC. Both png file and html file can be read from MS Word software to prepare a report of statistical analysis.

eStat allows both raw data type as in Figure 2 (a) and frequency data type as in Figure 2 (b) to draw a bar graph, a pie chart, and a band graph. A vertical bar graph by gender with using both type of data can be drawn as in Figure 2 (c) and it can be changed as side-by-side, stacked, ratio style, and horizontal style of a bar graph by using sub-graph icons.



File 021Discrete_MathPreferenc

Analysis Var 2: MathPref by Group 1: Sex

( Selected data: Raw Data ) (Summary Data: Mu

SelectedVar V2 by V1,

|  | Sex | MathPref | V3 | V4 |
|---|---|---|---|---|
| 1 | 1 | 3 |  |  |
| 2 | 2 | 1 |  |  |
| 3 | 1 | 3 |  |  |
| 4 | 2 | 1 |  |  |
| 5 | 1 | 3 |  |  |
| 6 | 1 | 1 |  |  |
| 7 | 1 | 2 |  |  |
| 8 | 2 | 2 |  |  |
| 9 | 2 | 3 |  |  |
| 10 | 1 | 2 |  |  |

File 000Summary_StudentBySex.csv

Analysis Var 1: Sex by Group 3: 5-2

( Selected data: Summary Data ) (Summary Data: Multip

SelectedVar V1 by V2,V3,

|  | Sex | 5-1 | 5-2 | V4 | V5 |
|---|---|---|---|---|---|
| 1 | Male | 16 | 12 |  |  |
| 2 | Female | 14 | 18 |  |  |

Figure 2. (a) Raw data, (b) Frequency data, (c) Bar graph by gender

‘eStatH’ icon at the main menu shows a separated menu at a new window for the use at both middle and high school students as in Figure 3. At the top of the menu, there are two eBook Links corresponding to Middle School and High School Statistics. If you click one of the links, the eBook will appear at a new window.

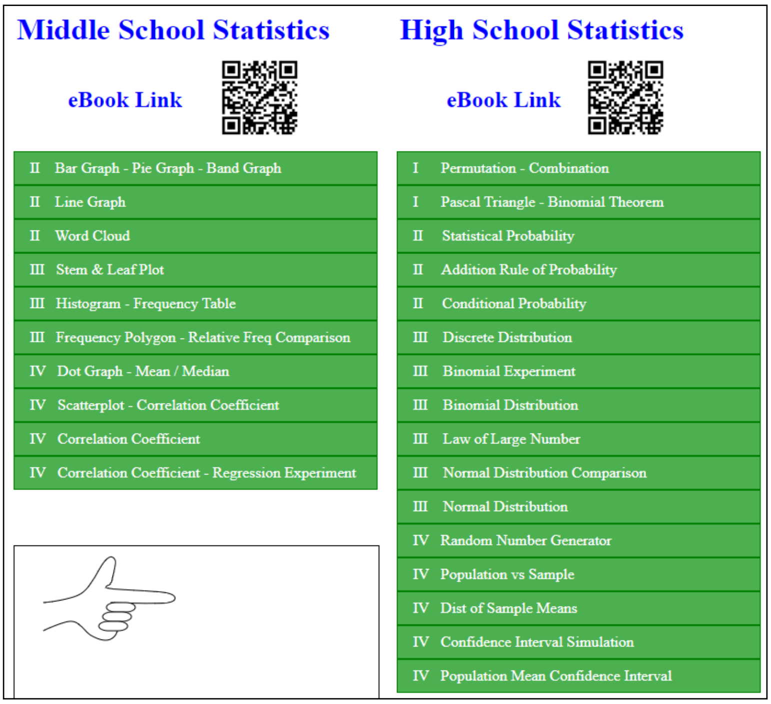

Middle School Statistics
eBook Link
II Bar Graph - Pie Graph - Band Graph
II Line Graph
II Word Cloud
III Stem & Leaf Plot
III Histogram - Frequency Table
III Frequency Polygon - Relative Freq Comparison
IV Dot Graph - Mean / Median
IV Scatterplot - Correlation Coefficient
IV Correlation Coefficient
IV Correlation Coefficient - Regression Experiment
High School Statistics
eBook Link
I Permutation - Combination
I Pascal Triangle - Binomial Theorem
II Statistical Probability
II Addition Rule of Probability
II Conditional Probability
III Discrete Distribution
III Binomial Experiment
III Binomial Distribution
III Law of Large Number
III Normal Distribution Comparison
III Normal Distribution
IV Random Number Generator
IV Population vs Sample
IV Dist of Sample Means
IV Confidence Interval Simulation
IV Population Mean Confidence Interval

Figure 3. (a) Menu for Middle School Statistics, (b) Menu for High School Statistics

'eStatU' icon at the main menu also shows a separated menu at a new window for the use of students at university as in Figure 4. If you click the eBook link at the top of the menu, the menu of eBook will appear at a new window as in Figure 5. This eBook can be also linked with 'eLearning' icon at the man menu. Modules of eStatU are linked with each topic of eBook to practice examples and study theories with various simulations.

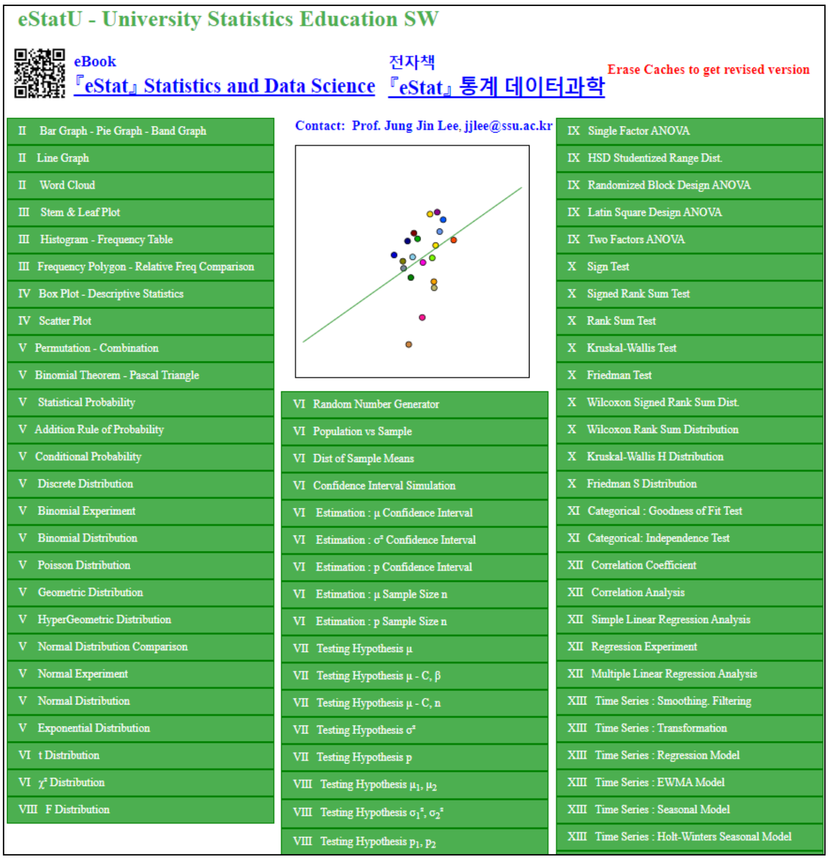
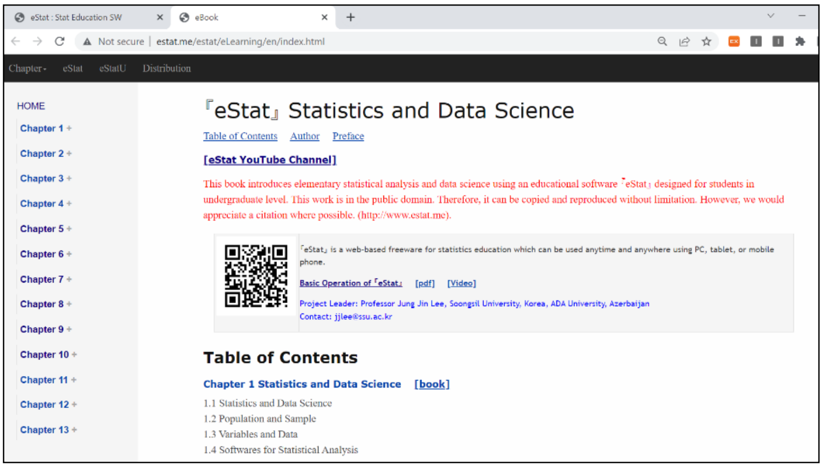

eStatU - University Statistics Education SW
eBook
『eStat』 Statistics and Data Science
전자책
『eStat』 통계 데이터과학
Erase Caches to get revised version
Contact: Prof. Jung Jin Lee, jjlee@ssu.ac.kr
Ⅱ Bar Graph - Pie Graph - Band Graph
Ⅱ Line Graph
Ⅱ Word Cloud
Ⅲ Stem & Leaf Plot
Ⅲ Histogram - Frequency Table
Ⅲ Frequency Polygon - Relative Freq Comparison
Ⅳ Box Plot - Descriptive Statistics
Ⅳ Scatter Plot
Ⅴ Permutation - Combination
Ⅴ Binomial Theorem - Pascal Triangle
Ⅴ Statistical Probability
Ⅴ Addition Rule of Probability
Ⅴ Conditional Probability
Ⅴ Discrete Distribution
Ⅴ Binomial Experiment
Ⅴ Binomial Distribution
Ⅴ Poisson Distribution
Ⅴ Geometric Distribution
Ⅴ HyperGeometric Distribution
Ⅴ Normal Distribution Comparison
Ⅴ Normal Experiment
Ⅴ Normal Distribution
Ⅴ Exponential Distribution
Ⅵ t Distribution
Ⅵ χ² Distribution
Ⅷ F Distribution
Ⅵ Random Number Generator
Ⅵ Population vs Sample
Ⅵ Dist of Sample Means
Ⅵ Confidence Interval Simulation
Ⅵ Estimation : μ Confidence Interval
Ⅵ Estimation : σ² Confidence Interval
Ⅵ Estimation : p Confidence Interval
Ⅵ Estimation : μ Sample Size n
Ⅵ Estimation : p Sample Size n
Ⅶ Testing Hypothesis μ
Ⅶ Testing Hypothesis μ - C, β
Ⅶ Testing Hypothesis μ - C, n
Ⅶ Testing Hypothesis σ²
Ⅶ Testing Hypothesis p
Ⅷ Testing Hypothesis μ₁, μ₂
Ⅷ Testing Hypothesis σ₁², σ₂²
Ⅷ Testing Hypothesis p₁, p₂
Ⅸ Single Factor ANOVA
Ⅸ HSD Studentized Range Dist.
Ⅸ Randomized Block Design ANOVA
Ⅸ Latin Square Design ANOVA
Ⅸ Two Factors ANOVA
Ⅹ Sign Test
Ⅹ Signed Rank Sum Test
Ⅹ Rank Sum Test
Ⅹ Kruskal-Wallis Test
Ⅹ Friedman Test
Ⅹ Wilcoxon Signed Rank Sum Dist.
Ⅹ Wilcoxon Rank Sum Distribution
Ⅹ Kruskal-Wallis H Distribution
Ⅹ Friedman S Distribution
Ⅺ Categorical : Goodness of Fit Test
Ⅺ Categorical: Independence Test
Ⅻ Correlation Coefficient
Ⅻ Correlation Analysis
Ⅻ Simple Linear Regression Analysis
Ⅻ Regression Experiment
Ⅻ Multiple Linear Regression Analysis
XIII Time Series : Smoothing. Filtering
XIII Time Series : Transformation
XIII Time Series : Regression Model
XIII Time Series : EWMA Model
XIII Time Series : Seasonal Model
XIII Time Series : Holt-Winters Seasonal Model

Figure 4. eStatU menu for university level of Statistics



『eStat』 Statistics and Data Science

Table of Contents Author Preface

[eStat YouTube Channel]

This book introduces elementary statistical analysis and data science using an educational software 『eStat』 designed for students in undergraduate level. This work is in the public domain. Therefore, it can be copied and reproduced without limitation. However, we would appreciate a citation where possible. (http://www.estat.me).

『eStat』 is a web-based freeware for statistics education which can be used anytime and anywhere using PC, tablet, or mobile phone.

Basic Operation of 『eStat』 [pdf] [Video]

Project Leader: Professor Jung Jin Lee, Soongsil University, Korea, ADA University, Azerbaijan
Contact: jjlee@ssu.ac.kr

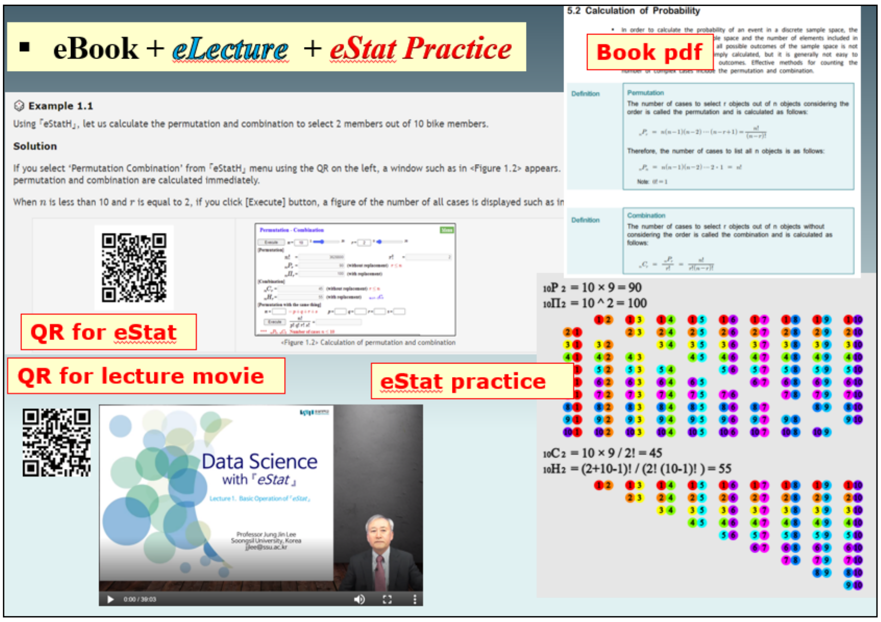
Table of Contents

Chapter 1 Statistics and Data Science [book]

1.1 Statistics and Data Science
1.2 Population and Sample
1.3 Variables and Data
1.4 Softwares for Statistical Analysis

Figure 5. eBook Menu for university level of Statistics

In addition, this eBook is linked with the pdf version of eBook, presentation file, and the lecture movie as in Figure 6. In section 3, useful features of the integrated system of eStat for Statistics education at undergraduate level are discussed.



Example 1.1

Using「eStatH」, let us calculate the permutation and combination to select 2 members out of 10 bike members.

Solution

If you select 'Permutation Combination' from「eStatH」menu using the QR on the left, a window such as in <Figure 1.2> appears. permutation and combination are calculated immediately.

When $n$ is less than 10 and $r$ is equal to 2, if you click [Execute] button, a figure of the number of all cases is displayed such as in

<Figure 1.2> Calculation of permutation and combination

5.2 Calculation of Probability

- In order to calculate the probability of an event in a discrete sample space, the ple space and the number of elements included in all possible outcomes of the sample space is not mply calculated, but it is generally not easy to outcomes. Effective methods for counting the number of complex cases include the permutation and combination.

Definition — Permutation

The number of cases to select r objects out of n objects considering the order is called the permutation and is calculated as follows:

$${}_nP_r = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

Therefore, the number of cases to list all n objects is as follows:

$${}_nP_n = n(n-1)(n-2)\cdots 2\cdot 1 = n!$$

Note: $0! = 1$

Definition — Combination

The number of cases to select r objects out of n objects without considering the order is called the combination and is calculated as follows:

$${}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n-r)!}$$

$${}_{10}P_2 = 10 \times 9 = 90$$

$${}_{10}\Pi_2 = 10 \wedge 2 = 100$$

$${}_{10}C_2 = 10 \times 9 / 2! = 45$$

$${}_{10}H_2 = (2+10\text{-}1)! / (2!\ (10\text{-}1)!\ ) = 55$$

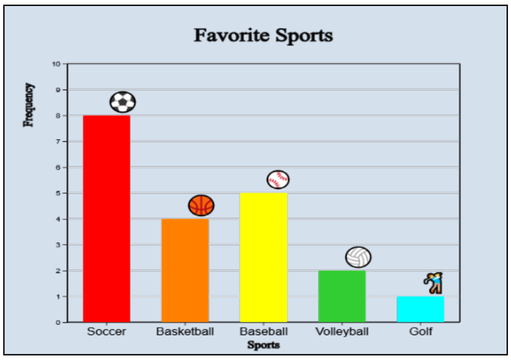
Figure 6. Integrated system of eStat which includes book, lecture movie and modules for practicing.

## 4. Useful features of eStat system

eStat includes modules for introductory statistical analysis which can be taught for two semesters at a university. Some useful features of eStat system for Statistics Education at university level which cannot be found in other statistical packages are introduced in this section.

*Data analysis by using dynamic graphs.*

Most of the graphs in eStat are dynamic by using the data driven document (d3) technology. Figure 7 (a) shows a colorful bar graph which is drawn dynamically from bottom to top of Graph Window. Particularly, bar graph, pie chart, band graph used for visualizing categorical data, and dot graph, histogram, stem-leaf plot, scatter plot used for visualizing quantitative data are drawn dynamically used for easily inviting school children or other beginners to the world of Statistics. Figure 7 (b) shows a colorful word cloud used for visualizing text data.



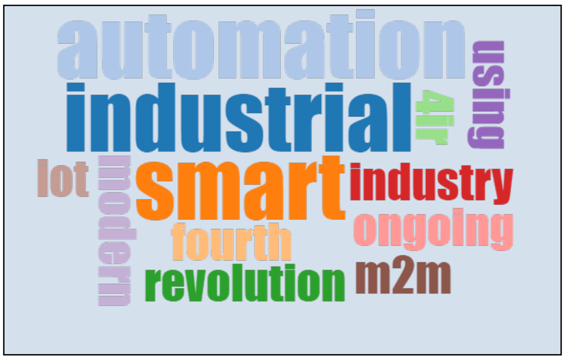
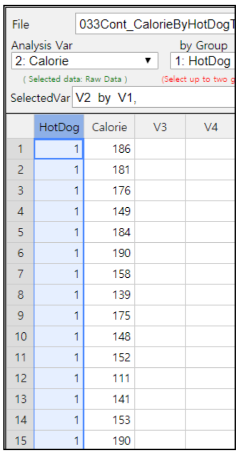


Figure 7. (a) Dynamic bar graph in eStat (b) Colorful word cloud in eStat

*Graphical output of Data Analysis*

Figure 8 (a) is a typical data for the analysis of variances which lists three types of hotdogs and their calories. Figure 8 (b) shows initial graphical analysis of data using dot graphs with means and also 95% confidence intervals based on each hotdog type. Figure 9 (a) shows the output of analysis of variance (ANOVA) with corresponding sampling distribution, F-distribution, value of test statistic, p-value, and Figure 9 (b) is a traditional output of ANOVA as provided by other statistical packages. A graph of the sampling distribution corresponding to each statistical analysis is drawn in a similar manner together with a traditional table output which is helpful to understand a difficult theory of Statistics. This graphical output of data analysis is the main usefulness of eStat for teaching Statistics unlike other traditional statistical packages.


File 033Cont_CalorieByHotDogT

Analysis Var 2: Calorie ▼ by Group 1: HotDog

( Selected data: Raw Data ) (Select up to two g

SelectedVar V2 by V1,

| | HotDog | Calorie | V3 | V4 |
|---|---|---|---|---|
| 1 | 1 | 186 | | |
| 2 | 1 | 181 | | |
| 3 | 1 | 176 | | |
| 4 | 1 | 149 | | |
| 5 | 1 | 184 | | |
| 6 | 1 | 190 | | |
| 7 | 1 | 158 | | |
| 8 | 1 | 139 | | |
| 9 | 1 | 175 | | |
| 10 | 1 | 148 | | |
| 11 | 1 | 152 | | |
| 12 | 1 | 111 | | |
| 13 | 1 | 141 | | |
| 14 | 1 | 153 | | |
| 15 | 1 | 190 | | |

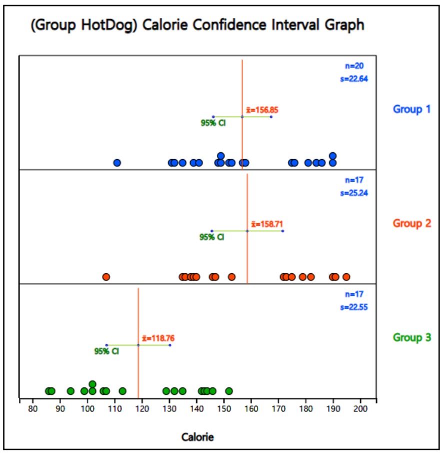
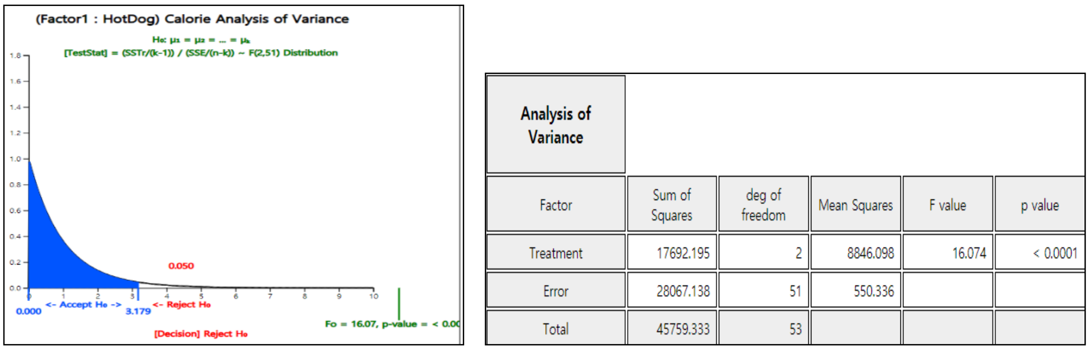


Figure 8. (a) Data for three types of hotdogs and their calories. (b) Graphical analysis with means and 95% confidence intervals

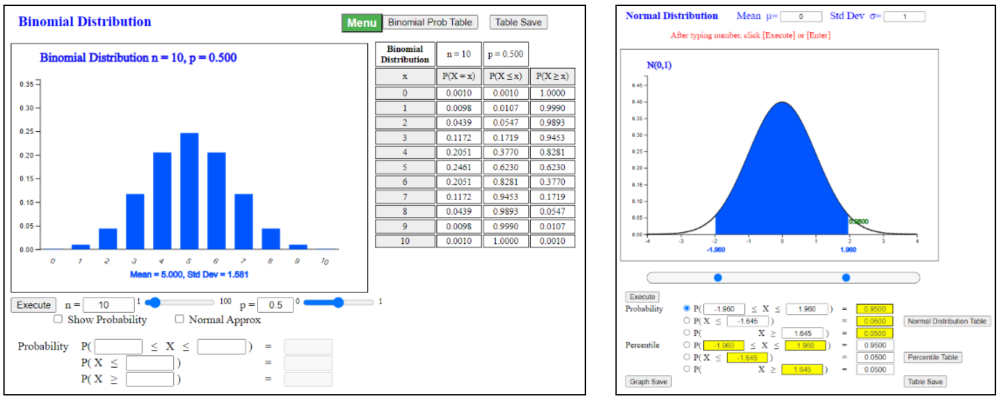

(Factor1 : HotDog) Calorie Analysis of Variance
H₀: μ₁ = μ₂ = ... = μₖ
[TestStat] = (SSTr/(k-1)) / (SSE/(n-k)) ~ F(2,51) Distribution
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
0.050
0.000
<- Accept H₀ ->
3.179
<- Reject H₀
Fo = 16.07, p-value = < 0.00
[Decision] Reject H₀

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Factor | Sum of Squares | deg of freedom | Mean Squares | F value | p value |
| Treatment | 17692.195 | 2 | 8846.098 | 16.074 | < 0.0001 |
| Error | 28067.138 | 51 | 550.336 | | |
| Total | 45759.333 | 53 | | | |

Figure 9. (a) Graphical output of ANOVA using F-distribution. (b) Traditional output of ANOVA using a table

*Tables and graphs of all statistical distributions.*

Traditional book of Statistics includes many tables of statistical distributions which occupy many pages. By using these tables, it is difficult to calculate a desired probability and a lecturer should spend a certain amount of time to teach how to use these tables. eStat includes a program of all statistical distributions which are commonly used in Statistics Education at a university. The program includes discrete distributions such as Binomial, Poisson, Geometric, Hypergeometric and continuous distributions such as Normal, Exponential, t, chi-square, F distribution. The program also includes non-parametric distributions such as Wilcoxon Signed Rank Sum, Wilcoxon Rank Sum, Kruskal-Wallis, Friedman distribution. eStat enables to easily show the graph of any distribution by changing values of parameters as illustrated in Figure 10 and calculate the probability of any interval type by using a dialog box. Therefore, eStat enables to show the graph of F sampling distribution and the p-value of ANOVA which were not possible in the traditional textbooks.



n = 10 p = 0.5

□ Show Probability □ Normal Approx

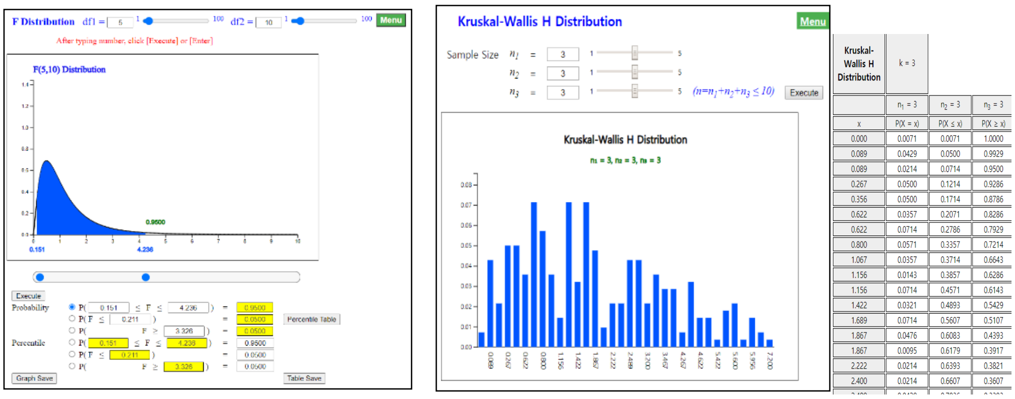
Probability P( ≤ X ≤ ) =
P( X ≤ ) =
P( X ≥ ) =

| Binomial Distribution | n = 10 | p = 0.500 | |
|---|---|---|---|
| x | P(X = x) | P(X ≤ x) | P(X ≥ x) |
| 0 | 0.0010 | 0.0010 | 1.0000 |
| 1 | 0.0098 | 0.0107 | 0.9990 |
| 2 | 0.0439 | 0.0547 | 0.9893 |
| 3 | 0.1172 | 0.1719 | 0.9453 |
| 4 | 0.2051 | 0.3770 | 0.8281 |
| 5 | 0.2461 | 0.6230 | 0.6230 |
| 6 | 0.2051 | 0.8281 | 0.3770 |
| 7 | 0.1172 | 0.9453 | 0.1719 |
| 8 | 0.0439 | 0.9893 | 0.0547 |
| 9 | 0.0098 | 0.9990 | 0.0107 |
| 10 | 0.0010 | 1.0000 | 0.0010 |

Mean μ= 0 Std Dev σ= 1

Probability ● P( -1.960 ≤ X ≤ 1.960 ) = 0.9500
○ P( X ≤ -1.645 ) = 0.0500
○ P( X ≥ 1.645 ) = 0.0500
Percentile ○ P( -1.960 ≤ X ≤ 1.960 ) = 0.9500
○ P( X ≤ -1.645 ) = 0.0500
○ P( X ≥ 1.645 ) = 0.0500

Figure 10. (a) Binomial distribution with table. (b) Normal distribution with probability calculation of an interval



df1 = 5 df2 = 10

Probability ● P( 0.151 ≤ F ≤ 4.236 ) = 0.9500
○ P( F ≤ 0.211 ) = 0.0500
○ P( F ≥ 3.326 ) = 0.0500
Percentile ○ P( 0.151 ≤ F ≤ 4.238 ) = 0.9500
○ P( F ≤ 0.211 ) = 0.0500
○ P( F ≥ 3.326 ) = 0.0500

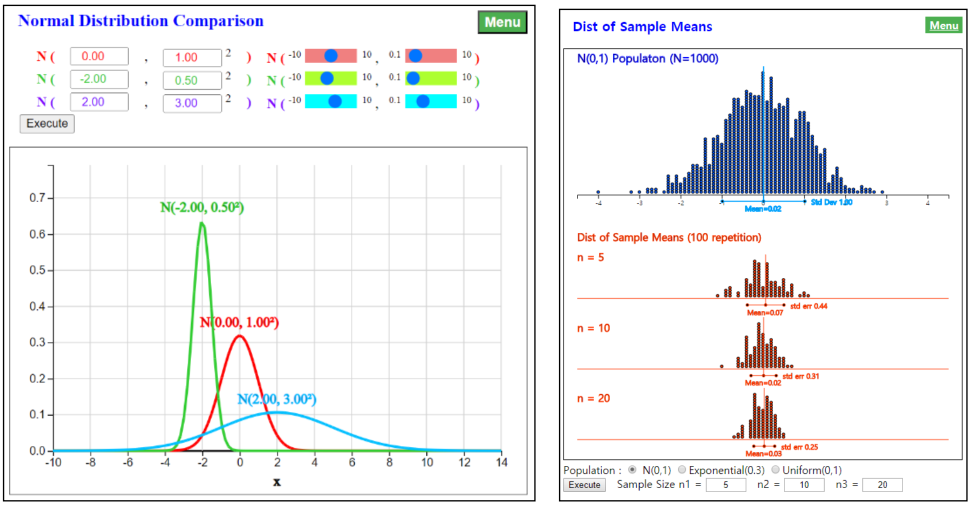
Sample Size $n_1$ = 3
$n_2$ = 3
$n_3$ = 3 $(n=n_1+n_2+n_3 \le 10)$

| Kruskal-Wallis H Distribution | k = 3 | | |
|---|---|---|---|
| | $n_1$ = 3 | $n_2$ = 3 | $n_3$ = 3 |
| x | P(X = x) | P(X ≤ x) | P(X ≥ x) |
| 0.000 | 0.0071 | 0.0071 | 1.0000 |
| 0.089 | 0.0429 | 0.0500 | 0.9929 |
| 0.089 | 0.0214 | 0.0714 | 0.9500 |
| 0.267 | 0.0500 | 0.1214 | 0.9286 |
| 0.356 | 0.0500 | 0.1714 | 0.8786 |
| 0.622 | 0.0357 | 0.2071 | 0.8286 |
| 0.622 | 0.0714 | 0.2786 | 0.7929 |
| 0.800 | 0.0571 | 0.3357 | 0.7214 |
| 1.067 | 0.0357 | 0.3714 | 0.6643 |
| 1.156 | 0.0143 | 0.3857 | 0.6286 |
| 1.156 | 0.0714 | 0.4571 | 0.6143 |
| 1.422 | 0.0321 | 0.4893 | 0.5429 |
| 1.689 | 0.0714 | 0.5607 | 0.5107 |
| 1.867 | 0.0476 | 0.6083 | 0.4393 |
| 1.867 | 0.0095 | 0.6179 | 0.3917 |
| 2.222 | 0.0214 | 0.6393 | 0.3821 |
| 2.400 | 0.0214 | 0.6607 | 0.3607 |
| [illegible] | [illegible] | [illegible] | [illegible] |

Figure 10. (c) F-distribution with probability calculation of an interval. (d) Kruskal-Wallis distribution with cumulative tail probability

*Simulation experiments to understand statistical theories.*

eStat includes many modules of simulation experiments to understand statistical theories. Figure 11 (a) is a module to compare shapes of Normal distributions depending on different types of parameters. Figure 11 (b) is a module to observe the implication of the Central Limit Theorem by using different sample sizes.



Normal Distribution Comparison

N( 0.00 , 1.00 $^2$ )

N( -2.00 , 0.50 $^2$ )

N( 2.00 , 3.00 $^2$ )

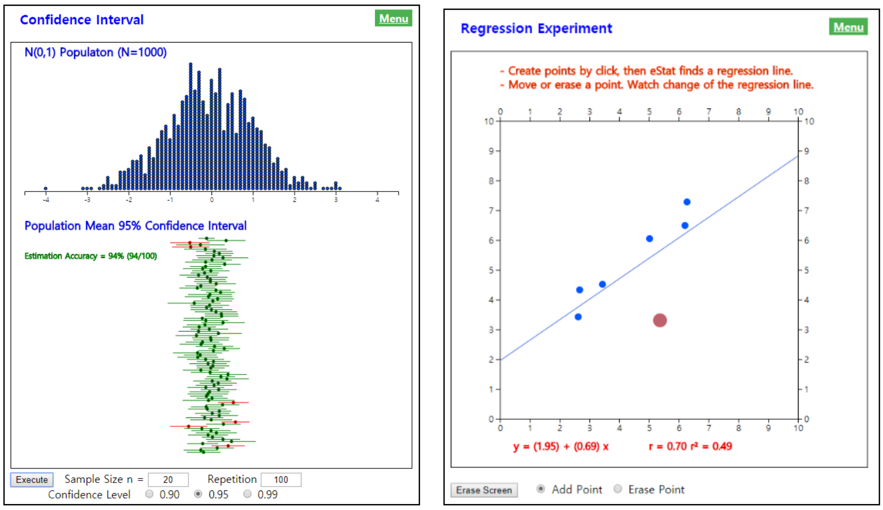
Dist of Sample Means

Population : ● N(0,1) ○ Exponential(0.3) ○ Uniform(0,1)

Sample Size n1 = 5 n2 = 10 n3 = 20

Figure 11. (a) Experiment for different parameters of Normal distribution. (b) Experiment for Central Limit Theorem.

Figure 12 (a) is a simulation module to watch the 95% confidence intervals by using different sample sizes and confidence levels. Figure 12 (b) is a simulation module to show the effect of an outlier in simple linear regression analysis by moving a single point.



Confidence Interval

Estimation Accuracy = 94% (94/100)

Sample Size n = 20 Repetition 100

Confidence Level ○ 0.90 ● 0.95 ○ 0.99

Regression Experiment

- Create points by click, then eStat finds a regression line.
- Move or erase a point. Watch change of the regression line.

y = (1.95) + (0.69) x r = 0.70 r² = 0.49

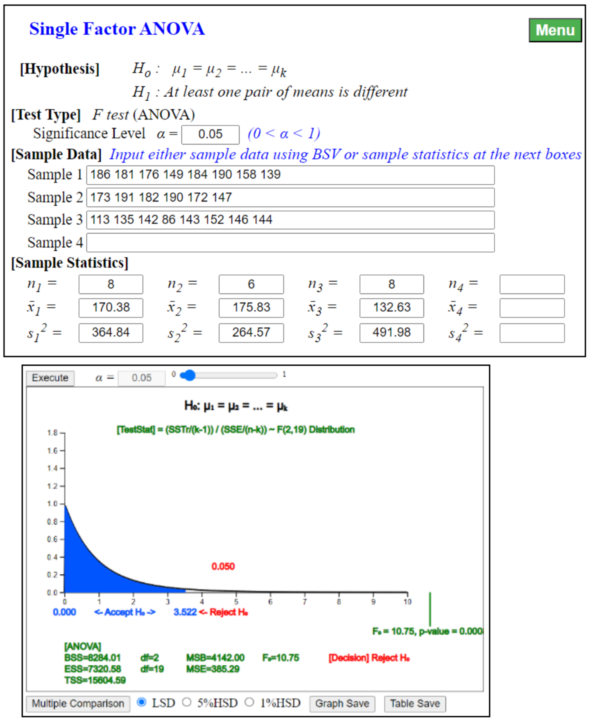
● Add Point ○ Erase Point

Menu
N(0,1) Populaton (N=1000)
Population Mean 95% Confidence Interval
Execute
Erase Screen
0
1
2
3
4
5
6
7
8
9
10

Figure 12. (a) Simulation of confidence interval. (b) Simulation of outlier in linear regression analysis.

*Intuitive design of data input in eStatU.*

Traditional data input of two columns, a factor variable and an analysis variable, for the analysis of variance as illustrated in Figure 8 (a) is not so familiar to beginners. eStatU has an intuitive design of data input for the analysis of variance as in Figure 13 (a) which immediately displays the sample statistics if you enter data and the graphical output with sampling distribution F as in Figure 13 (b). This design of data input in eStatU modules is useful to solve the exercise questions in the eBook.



**Single Factor ANOVA** Menu

[Hypothesis] $H_o: \mu_1 = \mu_2 = \ldots = \mu_k$

$H_1$ : *At least one pair of means is different*

[Test Type] *F test* (ANOVA)

Significance Level $\alpha =$ 0.05 $(0 < \alpha < 1)$

[Sample Data] *Input either sample data using BSV or sample statistics at the next boxes*

Sample 1 186 181 176 149 184 190 158 139

Sample 2 173 191 182 190 172 147

Sample 3 113 135 142 86 143 152 146 144

Sample 4

[Sample Statistics]

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $n_1 =$ | 8 | $n_2 =$ | 6 | $n_3 =$ | 8 | $n_4 =$ | |
| $\bar{x}_1 =$ | 170.38 | $\bar{x}_2 =$ | 175.83 | $\bar{x}_3 =$ | 132.63 | $\bar{x}_4 =$ | |
| $s_1^2 =$ | 364.84 | $s_2^2 =$ | 264.57 | $s_3^2 =$ | 491.98 | $s_4^2 =$ | |

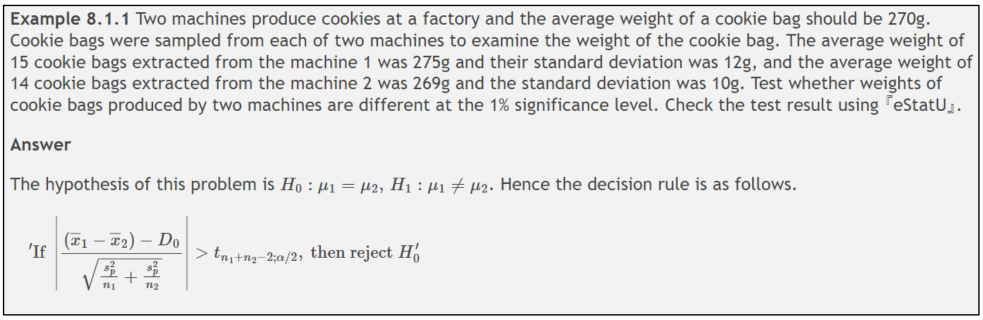
[ANOVA]
BSS=8284.01 df=2 MSB=4142.00 F₀=10.75 [Decision] Reject H₀
ESS=7320.58 df=19 MSE=385.29
TSS=15604.59

Multiple Comparison ● LSD ○ 5%HSD ○ 1%HSD Graph Save Table Save

Figure 13. (a) Data input of the analysis of variance. (b) Graphical output of the analysis of variance.

*Embedded modules in eBook to practice statistical analysis.*

Since eBook is also developed by HTML5, CSS3 and JavaScript, many of eStatU modules are embedded to each example of statistical analysis in eBook. Figure 14 shows an example of t-test for two population means with sample statistics only in eBook, and Figure 15 shows its data input and graphical output of t-test by using an embedded module of eStat. Students can practice any statistical analysis while they are studying an example with the embedded module. This embedded module has been useful for online teaching of Statistics during pandemic period in addition to the offline teaching.



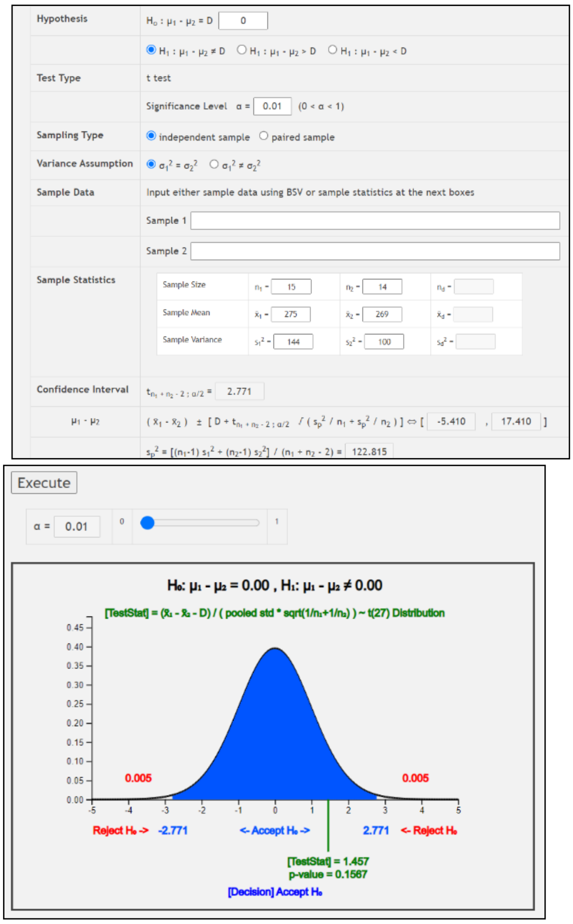
**Example 8.1.1** Two machines produce cookies at a factory and the average weight of a cookie bag should be 270g. Cookie bags were sampled from each of two machines to examine the weight of the cookie bag. The average weight of 15 cookie bags extracted from the machine 1 was 275g and their standard deviation was 12g, and the average weight of 14 cookie bags extracted from the machine 2 was 269g and the standard deviation was 10g. Test whether weights of cookie bags produced by two machines are different at the 1% significance level. Check the test result using 『eStatU』.

**Answer**

The hypothesis of this problem is $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$. Hence the decision rule is as follows.

$$\text{'If } \left| \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \right| > t_{n_1+n_2-2;\alpha/2}, \text{ then reject } H_0'$$

Figure 14. Example of t-test in eBook



| | |
|---|---|
| Hypothesis | $H_o : \mu_1 - \mu_2 = D$ 0 |
| | ● $H_1 : \mu_1 - \mu_2 \neq D$ ○ $H_1 : \mu_1 - \mu_2 > D$ ○ $H_1 : \mu_1 - \mu_2 < D$ |
| Test Type | t test |
| | Significance Level $\alpha =$ 0.01 $(0 < \alpha < 1)$ |
| Sampling Type | ● independent sample ○ paired sample |
| Variance Assumption | ● $\sigma_1^2 = \sigma_2^2$ ○ $\sigma_1^2 \neq \sigma_2^2$ |
| Sample Data | Input either sample data using BSV or sample statistics at the next boxes |
| | Sample 1 |
| | Sample 2 |
| Sample Statistics | |
| Confidence Interval | $t_{n_1+n_2-2\,;\,\alpha/2} =$ 2.771 |
| $\mu_1 - \mu_2$ | $(\bar{x}_1 - \bar{x}_2) \pm [D + t_{n_1+n_2-2\,;\,\alpha/2} \sqrt{(s_p^2 / n_1 + s_p^2 / n_2)}] \Leftrightarrow [$ -5.410 , 17.410 $]$ |
| | $s_p^2 = [(n_1-1)s_1^2 + (n_2-1)s_2^2] / (n_1 + n_2 - 2) =$ 122.815 |

| | | | |
|---|---|---|---|
| Sample Size | $n_1 =$ 15 | $n_2 =$ 14 | $n_d =$ |
| Sample Mean | $\bar{x}_1 =$ 275 | $\bar{x}_2 =$ 269 | $\bar{x}_d =$ |
| Sample Variance | $s_1^2 =$ 144 | $s_2^2 =$ 100 | $s_d^2 =$ |

Figure 15. (a) Data input of t-test with sample statistics only. (b) Graphical output of t-test with test statistic and p-value.

## 5. Conclusion

Since eStat modules are made by HTML5, JavaScript and D3.js, we encourage any developer to add his/her interesting idea, experience, and own modules. We hope that this free software eStat can help students at all levels of Statistics classroom around the world, especially developing countries. Consequently, many students become to realise the usefulness of Statistics in the era of information society.

## References

[1] CensusAtSchool, https://censusatschool.ie/about/

[2] Jung Jin Lee et al. (2019), Introduction to Data Science using eStat, Docuhut.

[3] Jung Jin Lee, Hyun Jo You (2018), Web Dynamic Data Visualization, Docuhut.

**[4] Jung Jin Lee, Tae Rim Lee, GunSeog Kang, Sungsoo Kim, Heon Jin Park, Yoon Dong Lee, Songyong Shim (2014), A Statistics Education Package Tong-Gramy for 5-8 Graders, The Korean Journal of Applied Statistics, 27(3), 421-429, 2014.**

**[5] TinkerPlots, https://en.wikipedia.org/wiki/TinkerPlots**

# A NEW PERSPECTIVE ON THE EVALUATION OF COMORBIDITY INDICES ON SURVIVAL

*Alev BAKIR[1], Benan MÜSELLİM[2], Mustafa Ş. ŞENOCAK[3]*

1 Istanbul University, Institute of Child Health, Department of Social Pediatrics,

alevbakir@yahoo.com

2 Istanbul University, Faculty of Cerrahpasa Medicine, Department of Chest Diseases (former lecturer)
benanmusellim@gmail.com

3 Istanbul University, Faculty of Cerrahpasa Medicine, Department of Biostatistics (former lecturer)
mssenocak@gmail.com

**Abstract**
Mortality studies must be evaluated by considering comorbid diseases. Existence, number, and type of comorbid diseases can have an important effect on prognosis. There are some approaches to estimate this affect with realistic and correct predictive power. Comorbid diseases' effect may differ according to comorbidity type and the main disease creating an important issue in survival analysis of patients with cancer. We aim to emphasize evaluation approaches giving different results in judgment results and prediction of prognosis. Retrospective cohort, data are collected from cases with Non-Small Cell Lung Cancer treated in Department of Chest Diseases 1998 to 2012. Their comorbid diseases' effects on the survival time are computed firstly univariately then according to number of comorbidities and finally all together by multivariate analysis and their specific combinations by considering some comorbid indices in the literature. It is crucial to take into comorbid diseases related to main disease and specially their combination when the risk is estimated in survival research. Reliable judgments of comorbid diseases list related to main disease have a great importance on development of this area.

***Keywords:*** *Comorbidity Index; Comorbid Diseases; Survival Analysis; Non-Small Cell Lung Cancer*

## 1. Introduction

Many types of research consider comorbid diseases significant, for instance survival not only dependent on pathologic stage, prognosis, age, and sex, but also on other factors such as comorbid diseases [1-2]. Additionally, comorbid diseases can affect the diagnosis, treatment, prognosis, and outcome [3].

In the literature, the effects of comorbid diseases are listed in various forms, such as scoring, severity of the comorbid diseases etc. [3]. Alvan Feinstein noted that "the failure to classify and analyse comorbid diseases has led to many difficulties in medical statistics" in the 1970s [4]. Previous comorbid indices approached more general to comorbid disease types, followed by age-adjusted or specific-disease comorbid indices [2-3, 5]. Comorbid indexes have been used frequently in studies on cancer, although there is no specific type of measurement or gold standard for cancer patients and comorbidity can play an important role in different types of research, and in some oncology studies it has a greater impact than age [6].

The Cumulative Illness Rating Scale (CIRS-1968), the Kaplan-Feinstein Classification (KFC-1974), the Charlson Comorbidity Index (CCI-1987) and the Index of Co-Existent Disease (ICED-1987) are valid and reliable and commonly used methods to measure comorbidity that can be used in clinical research [4]. Also, the most used is the CCI, the most detailed is the CIRS with scoring sheet, and the most complicated is ICED with scoring and physical condition, The KFC is a useful and realistic comorbidity index for clinical diabetes research because of specifically designed for diabetes [4, 6-7]. In addition to these, more current and specific indexes such as Modified Charlson Comorbidity Index, Elixhauser Comorbidity Measures, Ovarian Cancer Comorbidity Index (OCCI) are also available [6, 8-10]. These kinds of comorbid indices have been used regardless of the main disease however effect of comorbid diseases is changeable depending on the type of the main disease [9].

Comorbidity indexes are used in the studies or is tried to select the most suitable index for the study by comparing them, but the interaction of comorbid diseases was unobserved. The purpose of this study is to comorbid diseases' effects on the survival time is computed firstly univariately, then all together by multivariate analysis and finally according to number of comorbidities and their specific combinations by considering some comorbid indices in the literature. In this way we try to emphasize different evaluation approaches can give different results for both result judgment and prediction of prognosis in comorbid study. Our other aim is whether it is necessary new, original, and optimal indices by considering some comorbid indices in the literature.

## 2. Materials and Methods

### *2.1. Study population and data*

A retrospective cohort study was performed patient records from the Department of Chest Diseases 1998 to 2012. A homogeneous group was created from 455 cases by selecting 247 cases with non-small cell lung cancer (NSCLC) and no surgical operation. Data collected by file review included age, gender, type of treatment, survival, survival time and comorbid diseases. Comorbid diseases that were projected by senior consultant when selected are diabetes, chronic obstructive pulmonary disease (COPD), coronary heart disease, renal failure, asthma, interstitial lung disease, previously cancer. Sample size being insufficient by nature for reliable multivariate analysis, data is folded by four for more clearly statistical results when multivariate analysis.

### *2.2. Statistical analysis*

Descriptive statistics were used to describe the cases' characteristics and continuous data were expressed as mean, standard deviation (SD) and median throughout study. Categorical data were expressed as counts and proportions. Independent Student t Test or Mann Whitney U Test was used to test the difference between groups for independent two samples depending on normal distribution. The normal distribution of data was tested by Shapiro-Wilk Test. Survival Analysis was performed using the Kaplan-Meier method and survival difference between groups was tested using the Log-Rank Test [11-12]. Univariate and Multivariate Cox proportional hazard analysis within different time intervals determined risk factors for survival [13-15]. All data analysis was performed with SPSS (Statistical Package for the Social Sciences) 18 for windows and was reported with %95 confidence intervals, p<0.05 was considered significant.

## 3. Results

Of 247 non-small cell lung cancer (NSCLC) cases analysis, 220 (89%) were men and 27 (11%) were women. The mean age at time of diagnosis was 62.15±9.95 years (median:62), ranged from 34 to 87 years. Median duration of follow-up was 277 days, at the end of the follow-up 197 cases had died, 50 cases have still lived. Some cases have some comorbid diseases like diabetes, COPD, coronary heart disease, renal failure, asthma, interstitial lung disease, previously cancer. The highest rate of these comorbid diseases was coronary heart disease with 13% (n: 32) and the least rate was renal failure with 1% (n: 2). 172 cases did not have any comorbid disease therefore 25.5% of cases had just 1 comorbid disease, 3% of cases had 2 comorbid diseases and 2% of cases had 3 or more comorbid diseases.

No statistical difference was found in ages between male-female or died-alive (p= 0.096; 0.070, respectively). The difference between female and male for survival time was not statistically significant (514.04±637.59 median 247 and 461.20±554.47 median 287.50, respectively, p=0.809). The median survival time for NSCLC survival was 332 days (95% CI (confidence interval) 305.931-358.069) with Kaplan-Meier Analysis. Log-Rank Analysis didn't indicate a statistically significant difference in two genders' survival (p=0.529).

When evaluating risk factors for survival time with Cox proportional hazard analysis confirmed a statistically significant effect for age (p=0.002, Hazard Ratio (HR): 0.989; 95% CI:0.982-0.996) but didn't confirm for gender (p=0.209).

In univariate analysis, diabetes, COPD, renal failure and asthma comorbid diseases were not a statistically significant effect (p: 0.255; 0.317; 0.404; 0.337, respectively) but the following three comorbid diseases significantly affected survival time of cases: coronary heart disease (HR: 1.27; 95% CI:1.038-1.564) interstitial lung disease (HR: 12.29; 95% CI:7.308-20.652), previously cancer (HR: 1.62; 95% CI:1.131-2.306). In multivariate analysis, same comorbid diseases were still significantly effect on survival time and also they become more significant than univariate analysis, coronary heart disease (HR: 1.32; 95% CI:1.067-1.632), interstitial lung disease (HR: 13.17; 95% CI:7.816-22.187), previously cancer (HR: 1.68; 95% CI:1.174-2.402) (Table 1).

Table 1: Types of Comorbid Disease Separate Multivariate Analysis Results

| | β | S.E. | p | HR | 95% CI for HR |
|---|---|---|---|---|---|

| | | | | | Lower | Upper |
|---|---|---|---|---|---|---|
| Diabetes | 0.117 | 0.128 | 0.362 | 1.124 | 0.875 | 1.444 |
| COPD | 0.099 | 0.145 | 0.494 | 1.105 | 0.831 | 1.468 |
| Coronary Heart Disease | 0.277 | 0.108 | 0.010* | 1.320 | 1.067 | 1.632 |
| Renal Failure | -0.229 | 0.357 | 0.522 | 0.796 | 0.396 | 1.601 |
| Asthma | -0.333 | 0.260 | 0.201 | 0.717 | 0.430 | 1.194 |
| Interstitial Lung Disease | 2.578 | 0.266 | <0.001* | 13.168 | 7.816 | 22.187 |
| Previously Cancer | 0.518 | 0.183 | 0.005* | 1.679 | 1.174 | 2.402 |

* p<0.05 significant, S.E.: Standard Error, HR: Hazard Ratio; CI: Confidence Interval

Table 2 shows that analysis of total combinations in numbers; when cases who have two comorbid diseases whatever they are, compared with cases who haven't any comorbid disease (Reference Category), two comorbid diseases were significantly effect on survival time (HR: 1.80; 95% CI:1.225-2.640) and also founded same thing for cases who have four comorbid diseases whatever they are (HR: 9.94; 95% CI: 3.653-27.032).

Table 2: Number of Total Comorbid Diseases

| | | | | | | 95% CI for HR | |
|---|---|---|---|---|---|---|---|
| | n | β | S.E. | p | HR | Lower | Upper |
| 0 Comorbid Disease | 688 | | | <0.001* | | | |
| 1 Comorbid Disease | 252 | 0.142 | 0.084 | 0.091 | 1.152 | 0.978 | 1.357 |
| 2 Comorbid Diseases | 32 | 0.587 | 0.196 | 0.003* | 1.798 | 1.225 | 2.640 |
| 3 Comorbid Diseases | 12 | 0.122 | 0.292 | 0.675 | 1.130 | 0.637 | 2.005 |
| 4 Comorbid Diseases | 4 | 2.296 | 0.511 | <0.001* | 9.937 | 3.653 | 27.032 |

* p<0.05 significant, S.E.: Standard Error, HR: Hazard Ratio; CI: Confidence Interval

Even though table 2 is a general assessment, when we regard all types of comorbid combinations in analysis as it is expected more significant results and HR increase much more (Table 3). If we interpreted 2 comorbid diseases at random instead of content of combinations that have 2 diseases such as "diabetes+interstitial lung disease" or "diabetes+previously cancer", it might be inconvenient to show for comorbid diseases' effect on survival time. As is seen from Table 2, the HR of 2 comorbid diseases in cases was 1.80 but the HR of "diabetes+interstitial lung disease" combination comorbid disease was 59.52 as the HR of "diabetes+previously cancer" combination was 3.76 (Table 3). Although there was not a statistically significant difference for cases who have three comorbid diseases (Table 2), the HR of "diabetes+COPD+coronary heart disease" was 2.31 HR. In that case we face the risk of general assessment that takes number of total comorbid combination.

Table 3: Type of Comorbid Disease Combinations Multivariate Analysis Results

| | | | | | 95% CI for HR | |
|---|---|---|---|---|---|---|
| | β | S.E. | p | HR | Lower | Upper |
| Comorbid Combination | | | <0.001 | | | |
| Diabetes | -0.012 | 0.164 | 0.942 | 0.988 | 0.716 | 1.363 |
| COPD | -0.125 | 0.194 | 0.519 | 0.882 | 0.603 | 1.291 |
| Coronary Heart Disease | 0.264 | 0.123 | 0.033* | 1.302 | 1.022 | 1.658 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Renal Failure | -0.252 | 0.357 | 0.480 | 0.777 | 0.386 | 1.564 |
| Asthma | -0.009 | 0.292 | 0.976 | 0.991 | 0.559 | 1.758 |
| Interstitial Lung Disease | 2.380 | 0.303 | <0.001* | 10.803 | 5.970 | 19.546 |
| Previously Cancer | 0.241 | 0.229 | 0.292 | 1.273 | 0.813 | 1.994 |
| Diabetes+Coronary Heart Dis. | 0.138 | 0.357 | 0.698 | 1.149 | 0.570 | 2.314 |
| COPD+ Coronary Heart Dis. | 0.377 | 0.357 | 0.291 | 1.458 | 0.724 | 2.938 |
| Diabetes+Interstitial Lung Dis. | 4.086 | 0.550 | <0.001* | 59.517 | 20.262 | 174.823 |
| Diabetes+ Previously Cancer | 1.323 | 0.505 | 0.009* | 3.755 | 1.396 | 10.103 |
| COPD+ Previously Cancer | 0.843 | 0.504 | 0.094 | 2.323 | 0.866 | 6.233 |
| Diabetes + COPD + Coronary Heart Dis. | 0.838 | 0.357 | 0.019* | 2.312 | 1.148 | 4.658 |
| Diabetes + Coronary Heart Dis. + Asthma | -0.594 | 0.503 | 0.237 | 0.552 | 0.206 | 1.479 |
| Diabetes + COPD + Coronary Heart Dis. + Interstitial Lung Dis. | 2.412 | 0.512 | <0.001* | 11.158 | 4.093 | 30.420 |

* p<0.05 significant, S.E.: Standard Error, HR: Hazard Ratio; CI: Confidence Interval

When evaluating age and gender as risk factors for NCSLC survival time in all combined comorbid diseases, age was associated with 0.99 HR (95% CI:0.979-0.994) and combinations that were significant already are still significant and some combinations' HR were increase as before. For example, while "diabetes+interstitial lung disease" had 59.52 HR, it has become 70.55 HR.

Additionally, we applied other comorbid diseases to calculate for the most known indices that based on cases' fields with some potential limitation such as unrecorded severity of comorbid diseases. An interesting point, none of comorbid indices, Charlson Comorbidity Index, Kaplan-Feinstein Classification, Index of Co-Existent Disease and Cumulative Illness Rating Scale for Geriatrics, has significant effect for NSCLC on survival time (p: 0.684; 0.101; 0.273; 0.567, respectively).

## 4. Discussion

Researchers working with observational data require methods to appropriately adjust their analysis for underlying differences in cases' survival time [3-4, 16]. To improve the actual survival of cases, it is important to understand the impact of comorbidity on cases' survival [7, 17]. It can be said that some comorbid diseases' power was hide or increased by other diseases in the light of the foregoing findings, although there are some valid and reliable methods to measure effect of comorbidity that can be used in the literature [18-20]. As it is understood, taking just number of comorbid diseases when evaluating comorbidity is wrong and misleading.

In the light of such information:

- It is necessary to particularize real effects of comorbid diseases by doing together univariate analysis and multivariate analysis.
- Comorbid diseases should be taken together into consideration with their combinations not just single.
- Comorbidities that existing shouldn't be taken with number of total comorbid combination because different combination of comorbid diseases that have same number of diseases can have different power and significance, indices that have these approaches are untrustworthy.
- There aren't any indices that involve a comorbid disease list that is enough for all main diseases, it needs different indices because there are different comorbid diseases that be able to affect different main diseases' prognosis. Some comorbid indices may not include some comorbid diseases that have important effect for a main disease therefore using such indices isn't capable of an appropriate and confidential prediction.
- Indices that is formed number of comorbid disease list is untrustworthy and can't specify required details as a result of this estimations will be fallacious.

- Indices that assumed same power for every disease can be misleading. Indices should involve comorbid diseases not only weights between each other but also every disease's severity.

In conclusion, it is necessary to take account comorbid diseases in survival research, but they can give confidential results only if the evaluation is correct. Within the scope of importance of this property, doing research by clinicians for making lists of comorbid disease related to main disease and to index by taking severity of comorbid diseases into consideration have a great importance on development of this area.

## References

[1] Birim Ö., Kappetein A., Bogers Ad.J.J.C. Charlson Comorbidity ındex as a Predicto of Long-Term Outcome after Surgery for Nonsmall Cell Lung Cancer. Int Congr Ser. 2005; 28: 759-762.

[2] Groll D.L., To T., Bombardier C., Wright J.G. The Development of a Comorbidity Index with Physical Function as the Outcome. J Clin Epidemiol. 2005; 58: 595-602.

[3] Hall S.F. A User's Guide to Selection a Comorbidity İndex for Clinical Research. J Clin Epidemiol. 2006; 59: 849-855.

[4] Groot V., Beckerman H., Lankhorst G.J., Bouter L.M. How to Measure Comorbidity: a Critical Review of Avaliable Methods. J Clin Epidemiol. 2003; 56: 221-229.

[5] Tomoki Yamano, Shinichi Yamauchi, Kei Kimura, Akihito Babaya, Michiko Hamanaka, Masayoshi Kobayashi et al. Influence of age and comorbidity on prognosis and application of adjuvant chemotherapy in elderly Japanese patients with colorectal cancer:A retrospectivemulticentre study. European Journal of Cancer. 2017; 81: 90-101.

[6] Diana Sarfati. Review of methods used to measure comorbidity in cancer populations: No gold standard exists. Journal of Clinical Epidemiology. 2012; 65: 924-933.

[7] Extermann M. Measuring Comorbidity in Older Cancer Patients. Eur J Cancer. 2000; 36: 453-471.

[8] Licia Denti, Andrea Artoni, Monica Casella, Fabiola Giambanco, Umberto Scoditti and Gian Paolo Ceda. Validity of the Modified Charlson Comorbidity Index as Predictor of Short-term Outcome in Older Stroke Patients. Journal of Stroke and Cerebrovascular Diseases. 2015 (Feb); 24(2): 330-336.

[9] Anne Elixhauser, Claudias Teiner, D. Roberth Arris and Rosanna M. Coffey. Comorbidity Measures for Use with Administrative Data. Medical Care, 1998(Jan.); 36(1):8-27.

[10] Mette Calundann Noer, Cecilie Dyg Sperling, Sofie Leisby Antonsen, Bent Ottesen, Ib Jarle Christensen, Claus Høgdall. A new clinically applicable age-specific comorbidity index for preoperative risk assessment of ovarian cancer patients. Gynecologic Oncology. 2016; 141: 471–478.

[11] Şenocak M.Ş. Biyoistatistik ve Araştırma Yöntembilimi. İstanbul: İstanbul Tıp Kitabevi; 2014.

[12] Şenocak M.Ş. Özel Biyoistatistik: Epidemiyolojide Sayısal Çözümleme. İstanbul: Çağlayan Kitabevi; 1992.

[13] Kachigan SK. Multivariate Statistical Analysis. Radius Press; 1991.

[14] Kachigan SK. Statistical Analysis: An Interdisciplinary Introduction to Univariate&Multivariate Methods. Radius Press; 1986.

[15] Miller R. G. Survival Analysis. New York: John Wiley and Sons Inc.; 1998.

[16] Hyun-Ju Seo, Seok-Jun Yoon, Sang-Il Lee, Kun Sei Lee, Young Ho Yun, Eun-Jung Kim et al. A comparison of the Charlson comorbidity index derived from medical records and claims data from patients undergoing lung cancer surgery in Korea: a population-based investigation BMC Health Services Research 2010, 10:236.

[17] Larry B. Goldstein, Gregory P. Samsa, David B. Matchar and Ronnie D. Horner Charlson Index Comorbidity Adjustment for Ischemic Stroke Outcome Studies. Stroke. 2004 Aug; 35(8):1941-5.

[18] Nicolucci A, Cubasso D, Labbrozzi D, Mari E, Impicciatore P, Procaccini DA, et al. Effect of coexistent diseases on survival of patients undergoing dialysis. ASAIO J. 1992 Jul-Sep;38(3):M291-5.

[19] Mohamed L. Sorror, Michael B. Maris, Rainer Storb, Frederic Baron, Brenda M. Sandmaier, David G. Maloney et al. Hematopoietic cell transplantation (HCT)–specific comorbidity index: a new tool for risk assessment before allogeneic HCT. Blood. 2005 Oct 15; 106(8):2912-9.

[20] C. Hudon, M. Fortin, A. Vanasse. Cumulative Illness Rating Scale was a reliable and valid index in a family practice context. Journal of Clinical Epidemiology 58 (2005) 603–608.

# A NEW TWO-PARAMETER BIASED PREDICTION CLASS IN LINEAR MIXED MODELS

*Özge KURAN*[1],

[1]Dicle University, Faculty of Science, Department of Statistics, Diyarbakır/TURKEY

ozge.kuran@dicle.edu.tr;ozge-kuran@hotmail.com

**Abstract**

In this paper, we suggest a new two-parameter biased prediction class, will be called as modified Liu prediction, under multicollinearity in linear mixed models. By using the mean square error criterion, we give the necessary and sufficient conditions for the better performance of the new suggested modified Liu prediction over the ridge and Liu predictions defined already in linear mixed model literature. We handle the biasing parameters selection of modified Liu predictors. We complete the paper with a real data analysis to investigate the theoretical performance of our new proposed biased prediction class.

***Key Words:*** *The best linear unbiased prediction; Ridge prediction; Liu prediction; The modified Liu prediction; Multicollinearity*

## 1. Introduction

Linear mixed models (LMMs) [1] are extensively employed for analyzing data structures involving repeated measures, longitudinal, growth, and clustered (or nested) data. Unlike linear regression models (LRMs), LMMs incorporate both fixed effect parameters and random effect parameters. This enables LMMs to effectively capture and account for the variability and correlation present in the data.

LMMs are the following form

$$y = X\beta + Zu + \varepsilon \tag{1}$$

where $y$ is an $n \times 1$ vector of responses, $X$ is an $n \times p$ known design matrix for the fixed effects, $\beta$ is a $p \times 1$ parameter vector of fixed effects, $Z$ is an $n \times q$ known design matrix for the random effects, $u$ is a $q \times 1$ vector of random effects and $\varepsilon$ is an $n \times 1$ vector of random errors. It is assumed that $u$ and $\varepsilon$ follow independent and multivariate Gaussian distributions such that $\begin{bmatrix} u \\ \varepsilon \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}\right)$ where $G$ and $R$ are known positive definite (pd) matrices. Then, $var(y) = \sigma^2 H$ where $H = ZGZ' + R$. If $G$ and $R$ are unknown, we use maximum likelihood (ML) or restricted maximum likelihood (REML) estimation methods.

The best linear unbiased estimator (BLUE) ($\hat{\beta}$) and the best linear unbiased predictor (BLUP) ($\hat{u}$) are found by [2] and [3], respectively, as

$$\hat{\beta} = (X'H^{-1}X)^{-1}X'H^{-1}y$$

(2)

$$\hat{u} = GZ'H^{-1}(y - X\hat{\beta}). \tag{3}$$

In the real data world, it is quite natural that strong linear dependence arises between the columns of $X$ and this linear dependence situation is called as multicollinearity. Under multicollinearity case, we may encounter with some undesirable result like a large variance of $\hat{\beta}$ that deviates from its true value. To solve this undesirable result, estimators and predictors alternative to $\hat{\beta}$ and $\hat{u}$ can be suggested. The most commonly preferred approach to overcome multicollinearity problem is the ridge approach [4] in LRMs and by using [4], the ridge estimator ($\hat{\beta}_k$) and the ridge predictor ($\hat{u}_k$) are derived by [5] in LMMs, respectively, as

$$\hat{\beta}_k = \left(X'H^{-1}X + kI_p\right)^{-1}X'H^{-1}y$$

(4)

$$\hat{u}_k = GZ'H^{-1}\left(y - X\hat{\beta}_k\right)$$

(5)

where $k > 0$ is called as the ridge biasing parameter. $\hat{\beta}_k$ and $\hat{u}_k$ also examined by [6] with the help of Henderson's mixed model equations (MMEs) (see [3]).

Another popular attempt is Liu's approach [7] in LRMs and by using [7-9], the Liu estimator ($\hat{\beta}_d$) and the Liu predictor ($\hat{u}_d$) are derived by [10] in LMMs, respectively, as

$$\hat{\beta}_d = \left(X'H^{-1}X + I_p\right)^{-1}(X'H^{-1}y + d\hat{\beta})$$

(6)

$$\hat{u}_d = GZ'H^{-1}\left(y - X\hat{\beta}_d\right)$$

(7)

where $0 < d < 1$ is called as the Liu biasing parameter and $\hat{\beta}$ is the BLUE given by Eq. (2).

Both ridge and Liu predictors given by Eqs. (4-7) are biased prediction approaches that are suggested by using some prior information [8] in order to eliminate the negative effects of the multicollinearity problem in LMMs. In addition to two approaches, our goal in this article is to propose a new two-parameter biased prediction approach in LMMs by taking a convex combination of ridge and Liu predictors as prior information with the help of [11] in LRMs. This new approach called the modified Liu is such a convex combined approach that is resulted in unifying the advantages of ridge and Liu prediction approaches. Since it is a combination of both ridge and Liu approaches, it is thought to be more successful than ridge and Liu classes in minimizing the negative effects of multicollinearity. Then, the rest of this paper is structured as follows: Section 2, the new modified Liu estimator and predictor in LMMs are characterized. In Section 3, we make some mean square error comparisons and the biasing parameters selection of modified Liu predictors is handled in Section 4. A real data analysis is done in Section 5 and finally, discussion is given in Section 6.

## 2. New modified Liu prediction approach

The one parameter estimators and predictors such as ridge and Liu have a good manner of reducing the multicollinearity problem. But, it is better to combine the two biasing parameters of these one parameter estimators and predictors in describing new estimator and predictor. In addition to that, having two biasing parameters are more flexible and more efficient in getting a better estimator and predictor. Thus, by following [11], we will propose an estimator and predictor whose variance and bias are smaller than $\hat{\beta}_k$ and $\hat{\beta}_d$ where its length is expected to be closer to $\hat{\beta}$ than $\hat{\beta}_k$ and $\hat{\beta}_d$.

$u$ and $y$ are jointly Gaussian distributed as with the assumptions of model (1), $\begin{bmatrix} u \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ X\beta \end{bmatrix}, \sigma^2 \begin{bmatrix} G & GZ' \\ ZG & H \end{bmatrix}\right)$. And so, the conditional distribution of y given u is $y|u \sim N(X\beta + Zu, \sigma^2 R)$. The joint density of y and u (known as the MMEs) given by $f(y,u) = f(y|u)f(u)$

$$f(y,u) = (2\pi\sigma^2)^{-(n+q)/2}|R|^{-1/2}|G|^{-1/2} exp\left\{-\frac{1}{2\sigma^2}[(y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'G^{-1}u]\right\}$$

where $|.|$ denotes the determinant of a matrix. By dropping the constant term, the log-joint distribution of $f(y,u)$ is obtained as

$$logf(y,u) = logf(y|u) + logf(u) = -\frac{1}{2\sigma^2}\{[(y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'G^{-1}u]\}$$

and then, by following [11], we will minimize $logf(y,u)$ subject to $(\beta - (1-d)\hat{\beta})'(\beta - (1-d)\hat{\beta}) = c$ by adding the following penalization term with regularization parameter $\delta = -\frac{1}{2\sigma^2} \geq 0$ is added to $logf(y,u)$

$$logf(y,u) - \frac{1}{2\sigma^2}k[(\beta - (1-d)\hat{\beta})'(\beta - (1-d)\hat{\beta}) - c]. \qquad (8)$$

The partial derivatives of Eq. (8) with respect to the elements of $\beta$ and $u$ is taken and equals to zero and then, by writing $\hat{\beta}_{d,k}$ (the modified Liu estimator) and $\hat{u}_{d,k}$ (the modified Liu predictor) instead of $\hat{\beta}$ and $\hat{u}$

$$X'R^{-1}(y - X\hat{\beta}_{d,k}) + k(1-d)\hat{\beta} - \hat{\beta}_{d,k} - X'R^{-1}Z\hat{u}_{d,k} = 0 \qquad (9)$$

$$Z'R^{-1}(y - X\hat{\beta}_{d,k}) - (Z'R^{-1}Z + G^{-1})\hat{u}_{d,k} = 0. \qquad (10)$$

Eqs. (9) and (10) can also be written as the form of matrix as

$$\begin{pmatrix} X'R^{-1}X + I_p & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta}_{d,k} \\ \hat{u}_{d,k} \end{pmatrix} = \begin{pmatrix} (X'R^{-1}X + k(1-d)I_p)\hat{\beta} \\ Z'R^{-1}y \end{pmatrix}. \qquad (11)$$

Eq. (11) can be written as via [12] approach,

$$C\hat{\varphi} = \omega' R^{-1} y + \vartheta \tag{12}$$

where $\hat{\varphi} = (\hat{\beta}'_{d,k}, \hat{u}'_{d,k})'$, $\omega = (X, Z)$, $\vartheta = (k(1-d)\hat{\beta}', 0')'$, $C = \omega' R^{-1}\omega + G^{*+}$ with $G^* = \begin{bmatrix} I_p & 0 \\ 0 & G \end{bmatrix}$ and $G^{*+} = \begin{bmatrix} I_p & 0 \\ 0 & G^{-1} \end{bmatrix}$ where the superscript '+' denotes the Moore–Penrose inverse. $\hat{\varphi}$ is extract from Eq. (12) and $\hat{\varphi}$ is obtained as

$$\hat{\varphi} = C^{-1}\omega' R^{-1} y \; + \; C^{-1}\vartheta \tag{13}$$

where $C^{-1}$ is the inverse formula of the partitioned matrix (see [13]). $C^{-1}$ is substituted in Eq. (13) and then, the modified Liu estimator and predictor are found, respectively, as

$$\hat{\beta}_{d,k} = (X'H^{-1}X + I_p)^{-1}(X'H^{-1}X + k(1-d)I_p)\hat{\beta} \tag{14}$$

$$\hat{u}_{d,k} = GZ'H^{-1}\left(y - X\hat{\beta}_{d,k}\right). \tag{15}$$

The proposed new estimator and predictor given by Eqs. (14) and (15) are general estimator and predictor that include BLUE and Liu estimator:

1. If $k = (1/(1-d))$, $\hat{\beta}_{d,k}$ becomes $\hat{\beta}$.
2. If $k = (d/(1-d))$, $\hat{\beta}_{d,k}$ becomes $\hat{\beta}_d$.

## 3. Some mean square error comparisons

[14-16] studied $\mu = L'\beta + S'u$ for specific matrices $L \in \mathbb{R}^{p\times \dot{s}}$ and $S \in \mathbb{R}^{q\times \dot{s}}$ primarily for the case $\dot{s} = 1$. [2] derived the BLUP of $\mu$ for $\hat{\beta}$ and û as $\hat{\mu} = L'\hat{\beta} + S'\hat{u} = \mathbb{Q}\hat{\beta} + S'GZ'H^{-1}y$, where $\mathbb{Q} = L' - S'GZ'H^{-1}X$. Then, the ridge, Liu and modified Liu predictors of $\mu$ can be written as $\hat{\mu}_k = L'\hat{\beta}_k + S'\hat{u}_k = \mathbb{Q}\hat{\beta}_k + S'GZ'H^{-1}y$, $\hat{\mu}_d = L'\hat{\beta}_d + S'\hat{u}_d = \mathbb{Q}\hat{\beta}_d + S'GZ'H^{-1}y$ and $\hat{\mu}_{d,k} = L'\hat{\beta}_{d,k} + S'\hat{u}_{d,k} = \mathbb{Q}\hat{\beta}_{d,k} + S'GZ'H^{-1}y$.

The comparisons of $\hat{\mu}_{d,k}$ to $\hat{\mu}$, $\hat{\mu}_k$ and $\hat{\mu}_d$ are done according to the matrix mean square error (MMSE) criterion. Then, following [10] and [17], we obtain

$$MMSE(\hat{\mu}) = \mathbb{Q}MMSE(\hat{\beta})\mathbb{Q}' + \sigma^2 S'(G - GZ'H^{-1}ZG)S \tag{16}$$
$$MMSE(\hat{\mu}_k) = \mathbb{Q}MMSE(\hat{\beta}_k)\mathbb{Q}' + \sigma^2 S'(G - GZ'H^{-1}ZG)S \tag{17}$$
$$MMSE(\hat{\mu}_d) = \mathbb{Q}MMSE(\hat{\beta}_d)\mathbb{Q}' + \sigma^2 S'(G - GZ'H^{-1}ZG)S \tag{18}$$
$$MMSE(\hat{\mu}_{d,k}) = \mathbb{Q}MMSE(\hat{\beta}_{d,k})\mathbb{Q}' + \sigma^2 S'(G - GZ'H^{-1}ZG)S \tag{19}$$

where

$$MMSE(\hat{\beta}) = \sigma^2(X'H^{-1}X)^{-1} \tag{20}$$
$$MMSE(\hat{\beta}_k) = \sigma^2\left(X'H^{-1}X + kI_p\right)^{-1} + \mathrm{k}^2\left(\mathrm{X'H^{-1}X + kI_p}\right)^{-1}\beta\beta'\left(\mathrm{X'H^{-1}X + kI_p}\right)^{-1} \tag{21}$$

$$\begin{gathered}MMSE(\hat{\beta}_d) = \sigma^2\left(X'H^{-1}X + I_p\right)^{-1}(X'H^{-1}X + dI_p)\,(X'H^{-1}X)^{-1}(X'H^{-1}X + dI_p)\left(X'H^{-1}X + I_p\right)^{-1} + \\ (1-d)^2\left(X'H^{-1}X + I_p\right)^{-1}\beta\beta'\left(X'H^{-1}X + I_p\right)^{-1}\end{gathered} \tag{22}$$

$$\begin{gathered}MMSE(\hat{\beta}_{d,k}) = \sigma^2\left(X'H^{-1}X + I_p\right)^{-1}\left(X'H^{-1}X + k(1-d)I_p\right)(X'H^{-1}X)^{-1}\left(X'H^{-1}X + k(1- \\ d)I_p\right)\left(X'H^{-1}X + I_p\right)^{-1} + \left[\left(X'H^{-1}X + I_p\right)^{-1}\left(X'H^{-1}X + k(1-d)I_p\right) - I_p\right]\beta\beta'\left[\left(X'H^{-1}X + \\ I_p\right)^{-1}\left(X'H^{-1}X + k(1-d)I_p\right) - I_p\right].\end{gathered} \tag{23}$$

When we examine Eqs. (16), (17), (18) and (19), we can say that the superiority of $MMSE(\hat{\mu}_{d,k})$ over $MMSE(\hat{\mu})$, $MMSE(\hat{\mu}_k)$ and $MMSE(\hat{\mu}_d)$ is equivalent to the superiority of the *MMSE* values of $\hat{\beta}_{d,k}$ over $\hat{\beta}$, $\hat{\beta}_k$ and $\hat{\beta}_d$, respectively.

$MMSE(\hat{\mu}_{d,k})$, $MMSE(\hat{\mu}_k)$, $MMSE(\hat{\mu}_d)$ and $MMSE(\hat{\mu})$ are a little predigested if the transform model (1) to a canonical form via orthogonal transformation. Therefore, the use of the canonical form is preferred (see [10]). From Eqs. (20), (21), (22) and (23), we get

$$MMSE(\hat{\alpha}) = \sigma^2\Lambda^{-1} \tag{24}$$

$$MMSE(\hat{\alpha}_k) = \sigma^2\left(\Lambda + kI_p\right)^{-1}\Lambda\left(\Lambda + kI_p\right)^{-1} + k^2\left(\Lambda + kI_p\right)^{-1}\alpha\alpha'\left(\Lambda + kI_p\right)^{-1} \tag{25}$$

$$MMSE(\hat{\alpha}_d) = \sigma^2\left(\Lambda + I_p\right)^{-1}(\Lambda + dI_p)\Lambda^{-1}(\Lambda + dI_p)\left(\Lambda + I_p\right)^{-1} + (1-d)^2\left(\Lambda + I_p\right)^{-1}\alpha\alpha'\left(\Lambda + I_p\right)^{-1} \tag{26}$$

$$\begin{gathered}MMSE(\hat{\alpha}_{d,k}) = \sigma^2\left(\Lambda + I_p\right)^{-1}\left(\Lambda + k(1-d)I_p\right)\Lambda^{-1}\left(\Lambda + k(1-d)I_p\right)\left(\Lambda + I_p\right)^{-1} + \left[\left(\Lambda + I_p\right)^{-1}\left(\Lambda + \\ k(1-d)I_p\right) - I_p\right]\alpha\alpha'\left[\left(\Lambda + I_p\right)^{-1}\left(\Lambda + k(1-d)I_p\right) - I_p\right].\end{gathered} \tag{27}$$

where $\Lambda = diag(\lambda_i)$ the $p \times p$ orthogonal matrix of the eigenvalues of $X'H^{-1}X$ $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0)$.

**Theorem 3.1.** The estimator $\hat{\alpha}_{d,k}$ is superior to the estimator $\hat{\alpha}_k$ in the MMSE criterion iff

$$\begin{aligned}\alpha'\left(R_{d,k} - I_p\right)'&\left[\sigma^2\left((\Lambda + kI_p)^{-1}\Lambda(\Lambda + kI_p)^{-1} - R_{d,k}\Lambda^{-1}R_{d,k}\right)\right. \\ &\left. + \left((\Lambda + kI_p)^{-1}\Lambda - I_p\right)\alpha\alpha'\left((\Lambda + kI_p)^{-1}\Lambda - I_p\right)\right]^{-1}\left(R_{d,k} - I_p\right)\alpha < 1\end{aligned}$$

where $R_{d,k} = (\Lambda + I_p)^{-1}\left(\Lambda + k(1-d)I_p\right)$.

**Theorem 3.2.** The estimator $\hat{\alpha}_{d,k}$ is superior to the estimator $\hat{\alpha}_d$ in the MMSE criterion iff

$$\begin{aligned}\alpha'\left(R_{d,k} - I_p\right)'&\left[\sigma^2\left((\Lambda + I_p)^{-1}(\Lambda + dI_p)\Lambda^{-1}(\Lambda + dI_p)(\Lambda + I_p)^{-1} - R_{d,k}\Lambda^{-1}R_{d,k}\right)\right. \\ &\left. + \left((\Lambda + I_p)^{-1}(\Lambda + dI_p) - I_p\right)\alpha\alpha'\left((\Lambda + I_p)^{-1}(\Lambda + dI_p) - I_p\right)\right]^{-1}\left(R_{d,k} - I_p\right)\alpha < 1\,.\end{aligned}$$

For the proofs of Theorems 3.1 and 3.2, it can be seen [11] in LRMs. In this article, we have expanded these author's ideas in LRMs to the LMMs.

## 4. The biasing parameters selection

First, minimizing the $MMSE(\hat{\alpha}_{d,k})$ equation given by (27), then we get the optimal $k$ for a fixed value of $d$ as a follows

$$SMSE(\hat{\alpha}_{d,k}) = tr(MMSE(\hat{\alpha}_{d,k})) = \sigma^2\sum_{i=1}^{p}\frac{(\lambda_i + k(1-d))^2}{\lambda_i(\lambda_i + 1)^2} + \sum_{i=1}^{p}\frac{\alpha_i^2(k(1-d)-1)^2}{(\lambda_i + 1)^2} \tag{28}$$

where “tr” means “trace”. Differentiating according to $k$ and equals to zero, we have

$$\hat{k}_{mLiu} = \frac{1}{p}\sum_{i=1}^{p}\hat{k} \tag{29}$$

where $\hat{k} = \left(\frac{\lambda_i(\hat{\sigma}^2 - \hat{\alpha}_i^2)}{(d-1)(\lambda_i\hat{\alpha}_i^2 + \hat{\sigma}^2)}\right)$ and $\hat{\sigma}^2 = (y - X\hat{\beta})'H^{-1}(y - X\hat{\beta})/(n-p)$. Then, similarly, we get the optimal $d$ for a fixed value of $k$ by differentiating Eq. (28) according to $d$ and equals to zero, we have

$$\hat{d}_{mLiu} = \frac{1}{p}\sum_{i=1}^{p} \hat{d} \tag{30}$$

where $\hat{d} = \left(1 - \frac{\lambda_i(\hat{\alpha}_i^2 - \hat{\sigma}^2)}{\hat{k}(\lambda_i\hat{\alpha}_i^2 + \hat{\sigma}^2)}\right)$.

By expanding [11] in LRMs to LMMs, we determine the estimators of the biasing parameters $k$ and $d$ in an iterative method as follows:

1. Get the initial estimate of $d$ by $\hat{d} = \min_i \left(\frac{\hat{\alpha}_i^2}{\left(\frac{\hat{\sigma}^2}{\lambda_i}\right) + \hat{\alpha}_i^2}\right)$,
2. Compute $\hat{k}_{mLiu}$ from (29) using $\hat{d}$ in the point 1.
3. Compute $\hat{d}_{mLiu}$ in (30) using $\hat{k}_{mLiu}$ in the point 2.
4. If $\hat{d}_{mLiu} < 0$ or $\hat{d}_{mLiu} > 1$, use $\hat{d}_{mLiu} = \hat{d}$.

## 5. A real data analysis: the orthodontic growth data analysis

In this article, we utilize the orthodontic growth dataset originally introduced by [18]. Subsequently, [10] also employed this dataset for their research. The dataset comprises 27 children, with 16 males and 10 females. Researchers measured the distance (in millimeters) from the center of the pituitary to the pterygomaxillary fissure at ages 8, 10, 12, 14. The primary objective, as reported by [18], was to establish a model examining the relationship between the aforementioned distance and the age of children while considering gender differences. The plots displaying the distances against ages are available in [10] and [19]. The model applied to both male and female subjects $i$ $(i = 1, \dots, 27)$ at age $j$ $(j = 1, \dots, 3)$ can be defined as follows (refer to [19, p.343]):

$$distance_{ij} = \beta_0 + \beta_1 age_j + \beta_2 sex + \beta_3 age_j \times sex + u_i + \varepsilon_{ij}^{sex} \tag{31}$$

where the random effect $u_i$ is assumed to follow a normal distribution with mean 0 and variance $\sigma_u^2$ independent of the error term $\varepsilon_{ij}^{sex}$ which $N(0, \sigma_{\varepsilon sex}^2)$, where $\sigma_{\varepsilon sex}^2$ depends on sex. Let $1_q$, $0_q$ and $w_q$ respectively show $q \times 1$ vectors of ones, zeros and vector of repeated 8, 10, 12, 14 in order. Then, Eq. (32) can be written in the form of Eq. (1) with matrix notation where

$$y_{108\times1} = [distance_{1,1} \quad \dots \quad distance_{27,4}]' = [26.0 \quad \dots \quad 28.0]'$$

$$\mathrm{X}_{108\times4} = \begin{bmatrix} 1_{32} & \mathrm{w}_{32} & 0_{32} & 0_{32} \\ 1_{32} & \mathrm{w}_{32} & 0_{32} & 0_{32} \\ 1_{22} & \mathrm{w}_{22} & 1_{22} & \mathrm{w}_{22} \\ 1_{22} & \mathrm{w}_{22} & 1_{22} & \mathrm{w}_{22} \end{bmatrix}, \mathrm{Z}_{108\times27} = \begin{bmatrix} 0_{60} & 0_{16} & 0_4 & \cdots & 0_{52} \\ 1_4 & 1_4 & 1_4 & \ddots & 0_{52} \\ 0_{44} & 0_{88} & 0_{100} & \cdots & 1_4 \end{bmatrix}.$$

With the help of a hybrid optimization scheme used by R, the REML estimators of the variance parameters as $\widehat{\mathrm{G}}_{27\times27} = \widehat{\sigma}_{\mathrm{u}}^2 \times \mathrm{I}_{27} = 3.414 \times \mathrm{I}_{27}$ and

$$\widehat{\mathrm{R}}_{108\times108} = \widehat{\sigma}_{\varepsilon\mathrm{sex}}^2 \times \mathrm{I}_{108} = \begin{bmatrix} \widehat{\sigma}_{\varepsilon\mathrm{male}}^2 \times \mathrm{I}_{64} & 0_{44\times44} \\ 0_{44\times44} & \widehat{\sigma}_{\varepsilon\mathrm{female}}^2 \times \mathrm{I}_{44} \end{bmatrix} = \begin{bmatrix} 2.788 \times \mathrm{I}_{64} & 0_{44\times44} \\ 0_{44\times44} & 0.610 \times \mathrm{I}_{44} \end{bmatrix}.$$

The eigenvalues of the matrix $X'\widehat{H}^{-1}X$ are computed as$\lambda_1 = 2.6923 \times 10^{+3}$, $\lambda_2 = 0.4400 \times 10^{+3}$, $\lambda_3 = 0.0040 \times 10^{+3}$ and $\lambda_4 = 0.0005 \times 10^{+3}$. By calculating the condition number as $\frac{\lambda_{max}}{\lambda_{min}} = 5.0011 \times 10^{+3}$, we observe that it exceeds 1000. This indicates the presence of severe multicollinearity in the data.

Initial estimate of $d$ is calculated as $\hat{d} = 0.03480$ and then, we find $\hat{k} = \hat{k}_{mLiu} = 0.7262$ and $\hat{d}_{mLiu} = 0.03482$. If $\hat{d}_{mLiu} < 0$ or $\hat{d}_{mLiu} > 1$ condition isn't satisfied, we use $\hat{d} = \hat{d}_{mLiu} = 0.03482$.

**Table 1.** Parameter estimators and SMSE values

| | $\widehat{\boldsymbol{\beta}}_{k}$ | $\widehat{\boldsymbol{\beta}}_{d}$ | $\widehat{\boldsymbol{\beta}}_{d,k}$ |
|---|---|---|---|
| $\boldsymbol{\beta}_0$ | 15.9785 | 16.0293 | 14.5841 |
| $\boldsymbol{\beta}_1$ | 0.7565 | 0.7524 | 0.8603 |
| $\boldsymbol{\beta}_2$ | 2.3441 | 2.2810 | 2.9300 |
| $\boldsymbol{\beta}_3$ | -0.3219 | -0.3170 | -0.3725 |
| ***SMSE*** | 22.5754 | 28.5780 | **4.1553** |

**Table 2.** Parameter predictors

| | $\hat{\boldsymbol{u}}_{\boldsymbol{k}}$ | $\hat{\boldsymbol{u}}_{\boldsymbol{d}}$ | $\hat{\boldsymbol{u}}_{\boldsymbol{d,k}}$ |
|---|---|---|---|
| $\boldsymbol{u_1}$ | -1,0794 | -1,0843 | -0,8699 |
| $\boldsymbol{u_2}$ | -1,0794 | -1,0843 | -0,8699 |
| $\boldsymbol{u_3}$ | -0,7679 | -0,7729 | -0,5585 |
| $\boldsymbol{u_4}$ | -0,5603 | -0,5653 | -0,3509 |
| $\boldsymbol{u_5}$ | -0,4565 | -0,4615 | -0,2471 |
| $\boldsymbol{u_6}$ | -0,3527 | -0,3577 | -0,1433 |
| $\boldsymbol{u_7}$ | -0,0413 | -0,0462 | 0,1681 |
| $\boldsymbol{u_8}$ | -0,0413 | -0,0462 | 0,1681 |
| $\boldsymbol{u_9}$ | -0,3044 | -0,3085 | -0,1310 |
| $\boldsymbol{u_{10}}$ | 1,2675 | 1,2634 | 1,4409 |
| $\boldsymbol{u_{11}}$ | 0,6853 | 0,6803 | 0,8947 |
| $\boldsymbol{u_{12}}$ | 0,4777 | 0,4727 | 0,6871 |
| $\boldsymbol{u_{13}}$ | 1,7233 | 1,7184 | 1,9328 |
| $\boldsymbol{u_{14}}$ | 1,9310 | 1,9260 | 2,1404 |
| $\boldsymbol{u_{15}}$ | 2,8652 | 2,8603 | 3,0746 |
| $\boldsymbol{u_{16}}$ | 4,3185 | 4,3136 | 4,5279 |
| $\boldsymbol{u_{17}}$ | -4,4064 | -4,4035 | -4,1922 |
| $\boldsymbol{u_{18}}$ | -1,8924 | -1,8901 | -1,7197 |
| $\boldsymbol{u_{19}}$ | -1,8936 | -1,8907 | -1,6795 |
| $\boldsymbol{u_{20}}$ | -2,7997 | -2,8054 | -2,5582 |
| $\boldsymbol{u_{21}}$ | -0,7946 | -0,7923 | -0,6218 |
| $\boldsymbol{u_{22}}$ | 0,2601 | 0,2630 | 0,4742 |
| $\boldsymbol{u_{23}}$ | -1,3085 | -1,3086 | -1,1103 |
| $\boldsymbol{u_{24}}$ | -0,0049 | -0,0025 | 0,1678 |
| $\boldsymbol{u_{25}}$ | 1,3370 | 1,3399 | 1,5511 |
| $\boldsymbol{u_{26}}$ | 1,4075 | 1,4098 | 1,5803 |
| $\boldsymbol{u_{27}}$ | 3,1318 | 3,1347 | 3,3459 |

Based on the results from Table 1, we can deduce that the modified Liu estimator surpasses both the ridge and Liu estimators in terms of SMSE, with the chosen values of $\hat{k} = \hat{k}_{mLiu} = 0.7262$ and $\hat{d} = \hat{d}_{mLiu} = 0.03482$. This conclusion is further supported by the information presented in Figs. 1 and 2.

Considering Theorems 3.1 and 3.2, it can be observed that both theorems are satisfied. The computed conditions from Theorems 3.1 and 3.2 are $0.1019$ and $0.0991$, respectively, which are less than 1. This indicates that $\hat{\beta}_{d,k}$ outperforms $\hat{\beta}_k$ and $\hat{\beta}_d$ according to the MMSE criterion.

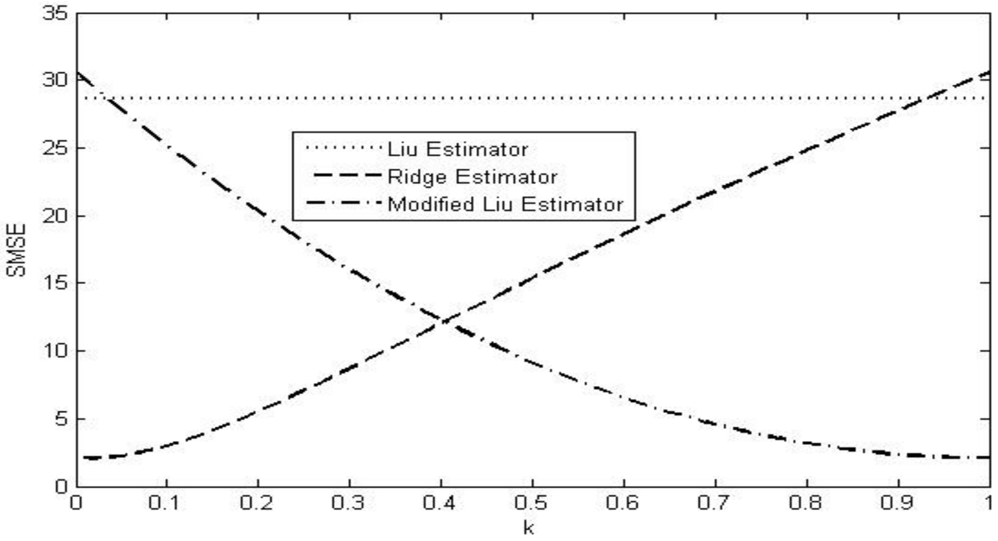


**Fig 1.** Plots of SMSE values of estimators versus k and fixed $\hat{d} = \hat{d}_{mLiu} = 0.03482$ values

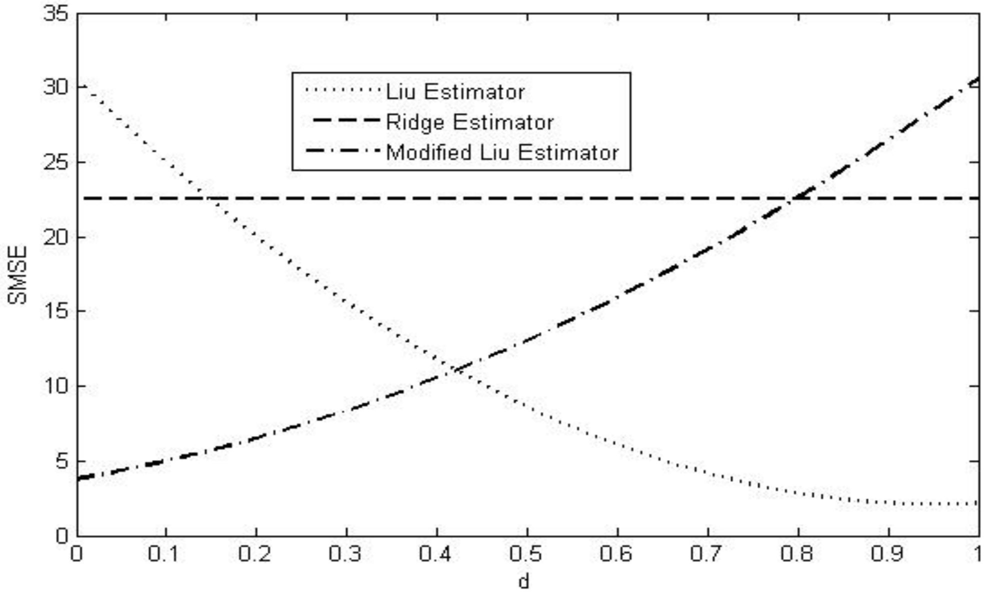


**Fig. 2.** Plots of SMSE values of estimators versus $d$ and fixed $\hat{k} = \hat{k}_{mLiu} = 0.7262$ values.

The analysis of orthodontic growth data suggests that there exist suitable values of $k$ and $d$ for which the modified Liu estimator outperforms both the ridge estimator and the Liu estimator, particularly when the condition of having small SMSE values is met.

## 6. Discussion

In the context of LMMs with multicollinearity, a new class of biased prediction with two parameters has been introduced. MMSE comparisons are conducted to determine the optimal biasing parameters. Real data analysis is performed to demonstrate the findings. The results of the analysis reveal that the superiority of the $\hat{\beta}_{d,k}$ estimator depends on the chosen biasing parameter values. When the appropriate biasing parameter condition is met, the modified Liu estimator exhibits a smaller SMSE value compared to the ridge and Liu estimators.

## References

[1] Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982;**38**:963 – 74.

[2] Henderson CR. Estimation of genetic parameters (abstract). *Ann Math Statis* 1950;**21**:309 – 10.

[3] Henderson CR, Kempthorne O, Searle SR. Estimation of environmental and genetic trends from records subject to culling. *Biometrics* 1959;**15**:192 – 218.

[4] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970;**12**:55 – 67.

[5] Liu XQ, Hu P. General ridge predictors in a mixed linear model. *J Theor Appl Statist* 2013;**47**:363 – 78.

[6] Özkale MR, Can F. An evaluation of ridge estimator in linear mixed models: an example from kidney failure data. *J Appl Statist* 2017;**44**:2251 – 69.

[7] Liu K. A new class of biased estimate in linear regression. *Commun Statist Theory Methods* 1993;**22**:393–402.

[8] Swindel BF. Good estimators based on prior information. *Commun Statist Theory Methods* 1976;**5**:1065–75.

[9] Özkale MR, Kaçıranlar S. The restricted and unrestricted two-parameter estimators. *Commun Statist Theory Methods* 2007;**36**:2707–25.

[10] Özkale MR, Kuran Ö. A further prediction method in linear mixed models: Liu prediction. *Comm Statist Simul Comput* 2020;**49**:3171 – 95.

[11] Dawoud I, Abonazel MR, Awwad FA. Modified Liu estimator to address the multicollinearity problem in regression models: A new biased estimation class. *Sci Afr* 2022;**17**:e01372.

[12] Gilmour AR, Thompson R, Cullis BR. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 1995;**51**:1440 – 50.

[13] Searle SR. *Matrix algebra useful for statistics*. New York: JohnWiley and Sons; 1982.

[14] Yang H, Ye H, Xue K. A further study of predictions in linear mixed models. *Commun Statist Theory Methods* 2014;**43**:4241–52.

[15] Pereira LN, Coelho PS. A small area predictor under area-level linear mixed models with restrictions. *Commun Statist Theory Methods* 2012;**41**:2524-44.

[16] Robinson GK. That BLUP is a good thing: the estimation of random effects (with discussion). *Stat Sci* 1991;**6**:15-51.

[17] Štulajter F. Predictions in nonlinear regression models. *Acta Math Univ Comen* 1997;**LXVI**:71–81.

[18] Potthoff RF, Roy SN. A generalized multivariate analysis of variance model especially useful for growth curve problems. *Biometrika* 1964;**51**:313–26.

[19] Sheather SJ. *A modern approach to regression with R*. New York: Springer Science and Business Media; 2009.