

Introduction to Data Science and Artificial Intelligence using 『eStat』 and R

This book introduces data science and artificial intelligence using 『eStat』 and R.

[Author](#) [Preface](#)

Project leader: Professor Jung Jin Lee, email: jjlee@ssu.ac.kr
Soongsil University, Korea, ADA University, Azerbaijan
New Uzbekistan University, Uzbekistan

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』 \(/estat/eLearning/en/eStatBasicOperation.html\)](#)

[\[pdf\] \(/estat/eLearning/en/eStatBasicOperation.pdf\)](#) [\[Video\]](#)

<https://www.youtube.com/watch?v=jjfCqdDZZ6Q>



R is a free software environment for statistical computing and graphics.

R site and to download; <https://www.r-project.org> (<https://www.r-project.org/>)

[Basic operation of R \[pdf\] \(/estat/eLearning/en/DataScience/rBasicOperation.pdf\)](#)

Table of Contents [\[book\] \(./pdf/book0.pdf\)](#)

[Chapter 1 Data science and artificial intelligence \(chapter01.html\)](#)

[\[book\] \(./pdf/book1.pdf\)](#)

- 1.1 Statistics, data science, machine learning, and artificial intelligence
- 1.2 General process of data analysis
- 1.3 Data classification
- 1.4 Software programs for data analysis
- 1.5 References

[Chapter 2 Data visualization \(chapter02.html\)](#) [\[book\] \(./pdf/book2.pdf\)](#)

- 2.1 Visualization of qualitative data
- 2.2 Visualization of quantitative data

2.3 R practice

Chapter 3 Data summary and transformation (chapter03.html) **[book]** **(./pdf/book3.pdf)**

- 3.1 Categorical data summary using tables
- 3.2 Quantitative data summary using measures
- 3.3 Data manipulation and transformation
- 3.4 Dimension reduction: Principal component analysis
- 3.5 R practice

Chapter 4 Probability and distribution (chapter04.html) **[book]** **(./pdf/book4.pdf)**

- 4.1 Probability
- 4.2 Random variable and distribution
- 4.3 Multivariate probability distribution
- 4.4 Estimation of a distribution

Chapter 5 Testing hypothesis and regression (chapter05.html) **[book]** **(./pdf/book5.pdf)**

- 5.1 Sampling distribution and estimation
- 5.2 Testing hypothesis for a population mean
- 5.3 Testing hypothesis for two populations meanss
- 5.4 Testing hypothesis for several population means: Analysis of variance
- 5.5 Regression analysis
- 5.6 R practice

Chapter 6 Supervised machine learning for categorical data (chapter06.html) **[book]** **(./pdf/book6.pdf)**

- 6.1 Basic concepts of supervised machine learning and classification
- 6.2 Decision tree model
- 6.3 Naive Bayes classification model
- 6.4 Evaluation and comparison of classification model

Chapter 7 Supervised machine learning for continuous data (chapter07.html) **[book]** **(./pdf/book7.pdf)**

- 7.1 Bayes classification model
- 7.2 Logistic regression model
- 7.3 Nearest neighbor classification model
- 7.4 Neural network model
- 7.5 Support vector machine model
- 7.6 Ensemble model
- 7.7 Classification of multiple groups

Chapter 8 Unsupervised machine learning (chapter08.html) **[book]** **(./pdf/book8.pdf)**

8.1 Basic concepts of unsupervised machine learning and clustering

8.2 Hierarchical clustering model

8.3 K-Means clustering model

Chapter 9 Artificial intelligence and other applications (chapter09.html). [book] (./pdf//book9.pdf)

9.1 Artificial intelligence, machine learning, and deep learning

9.2 Text mining

9.3 Web data mining

9.4 Multimedia data mining

9.5 Spatial data analysis

Authors of eBook and developers of 『eStat』

Jung Jin Lee

Emeritus Professor, Soongsil University, Korea

Professor, ADA University, Azerbaijan

B.S. and M.S., Seoul National University

Ph.D. in O.R., Case Western Reserve University

President, Korean Statistical Society

Vice President, International Association for Statistical Computing

Council Member, International Statistical Institute (ISI)



Tae Rim Lee

Emeritus Professor, Korea National Open University

B.S. and M.S., Seoul National University

Ph.D. in Statistics, Choongang University

Vice President, Korean Statistical Society

Vice President, International Association for Statistics Education

Vice President, International Biometric Society



Geunseog Kang

Emeritus Professor, Soongsil University, Korea

B.S. and M.S., Seoul National University

Ph.D. in Statistics, University of Wisconsin - Madison



Sung Soo Kim

Professor, Korea National Open University

B.S., M.S., Ph.D. in Statistics, Seoul National University



Heon Jin Park

Professor, Inha University
 B.S. and M.S., Seoul National University
 Ph.D. in Statistics, Iowa Stat University
 President, Korean Data Mining Society

**Song Yong Sim**

Hallym University
 B.S. and M.S., Seoul National University
 Ph.D. in Statistics, University of Wisconsin - Madison

**Yoon Dong Lee**

Professor, Sogang University
 B.S. and M.S., Seoul National University
 Ph.D. in Statistics, Iowa State University

**Hyun Jo You**

Professor, Chungnam National University
 B.S., M.S., Ph.D in Linguistics, Seoul National University
 Ph.D in Statistics, Soongsil University



Preface

Over the last half century, Computer Science has been evolved at a tremendous rate, bringing about previously unimaginable changes in many areas of our society and enriching our lives. Recent merging of Computer Sciences with Communication Technologies has created a digital revolution called the 4th industrial revolution that will lead for another future change.

The 4th Industrial Revolution aims at super-connectedness, super-intelligence and super-forecasting and many new changes will occur in our lives revolutionary. The revolution would help us to solve many problems, but it would also give us new challenges to be solved at the same time. The biggest challenge is analysis and utilization of Big Data.

The analysis of Big Data can be done by multi-disciplinary areas such as Statistics, Mathematics, Computer Science, and other application areas such as Management which is called Data Science. Data Science is primarily based on traditional statistical methods, applied mathematics, and requires lots of data manipulation using computer software such as R, SAS and SPSS which are widely used require some training from professionals. Authors of this book have been developed 『eStat』 for years which can help all level of students to learn Data Science easily.

This book introduces basic visualization data in Chapter 2, data summary and transformation in Chapter 3. Chapter 4 and 5 review basic statistical model for big data analysis. Chapter 6 and 7 discuss models of supervised machine learning, and Chapter 8 discusses models of unsupervised machine learning. Chapter 9 introduces artificial intelligence and other applications of data science.

I appreciate all of you who have developed 『eStat』 together over the past few years. I appreciate also to all internet communities who have helped us during the development of 『eStat』.

Spring 2025

Project Leader: Jung Jin Lee