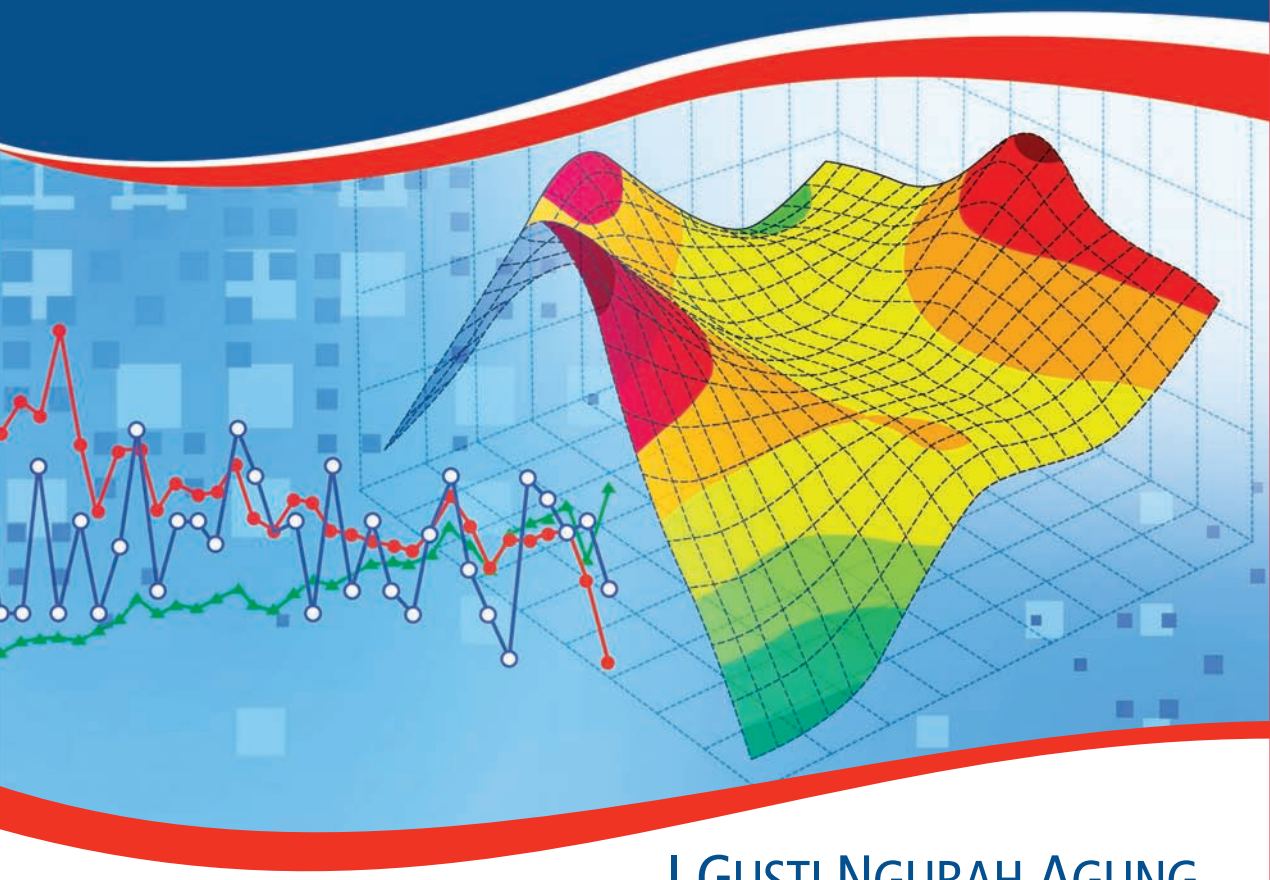
# Cross Section and Experimental Data Analysis Using Eviews



I Gusti Ngurah Agung

WILEY

# CROSS SECTION AND EXPERIMENTAL DATA ANALYSIS USING EVIEWS

# CROSS SECTION AND EXPERIMENTAL DATA ANALYSIS USING EVIEWS

**I Gusti Ngurah Agung**
*Graduate School of Management*
*Faculty of Economics and Business, University of Indonesia, Indonesia*

*Ph.D. in Biostatistics and*
*MSc. in Mathematical Statistics from*
*University of North Carolina at Chapel Hill*

John Wiley & Sons (Asia) Pte Ltd

This edition first published 2011


***Registered office***
John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop, # 02-01, Singapore 129809



***Library of Congress Cataloging-in-Publication Data***

Agung, I Gusti Ngurah.
Cross section and experimental data analysis using EViews / I Gusti Ngurah Agung.
p. cm.
ISBN 978-0-470-82842-7 (cloth)
1. Statistics. 2. EViews (Computer file) I. Title.
HA29.A376 2011
005.5'5–dc22 2010041053

Print ISBN: 978-0-470-82842-7
ePDF ISBN: 978-0-470-82843-4
oBook ISBN: 978-0-470-82844-1
ePub ISBN: 978-0-470-82845-8

Set in 10/12pt Times by Thomson Digital, Noida, India.

*Dedicated to my wife*
*Anak Agung Alit Mas,*
*my children*
*Martiningsih A. Chandra, Ratnaningsih A. Lefort, and Darma Putra,*
*my sons in law*
*Aditiawan Chandra, and Eric Lefort,*
*my daughter in law*
*Refiana Andries, and*
*all my grandchildren*
*Indra, Rama, Luana, Leonard,*
*Agung Mas Mirah, and Agung Surya Buana*

# Contents

**Preface** **xv**

**1 Misinterpretation of Selected Theoretical Concepts of Statistics** **1**
1.1 Introduction 1
1.2 What is a Population? 2
1.3 A Sample and Sample Space 2
1.3.1 What is a Sample? 2
1.3.2 What is the Sample Space? 3
1.3.3 What is a Representative Sample? 6
1.3.4 Relationship between the Sample Space, Population, and a Sample 7
1.4 Distribution of a Random Sample Space 8
1.5 What is a Random Variable? 9
1.6 Theoretical Concept of a Random Sample 9
1.6.1 What is a Random Sample in Statistics? 9
1.6.2 Central Limit Theorem 10
1.6.3 Unbiased Statistics based on Random Samples 16
1.6.4 Special Notes on Nonrandom Sample 19
1.7 Does a Representative Sample Really Exist? 19
1.8 Remarks on Statistical Powers and Sample Sizes 21
1.9 Hypothesis and Hypothesis Testing 24
1.10 Groups of Research Variables 25
1.10.1 Problem Indicators 26
1.10.2 Controllable Cause Factors 26
1.10.3 Uncontrollable Cause Factors 26
1.10.4 Background or Classification Factors 27
1.10.5 Environmental Factors 27
1.11 Causal Relationship between Variables 27
1.11.1 Bivariate Correlation 27
1.11.2 Special Remarks 30
1.12 Misinterpretation of Selected Statistics 31
1.12.1 Standard Error 31
1.12.2 Significance Level and Power of a Test 31
1.12.3 Reliability of a Test or Instrument 32

1.12.4 Validity of a Test or Instrument 33
1.12.5 Reliability and Validity of Forecasting 34
1.12.6 Reliability and Validity of a Predicted Risk 35

**2 Simple Statistical Analysis but Good for Strategic Decision Making 37**
2.1 Introduction 37
2.2 A Single Input for Decision Making 39
2.2.1 A Single Sampled Unit 39
2.2.2 Descriptive Statistics Based on a Single Measurable Variable 39
2.2.3 Agung Six-Point Scale (ASPS) Problem Indicator 43
2.2.4 Latent Variables and Composite Indexes 45
2.2.5 Demographic and Social–Economic Factors 45
2.2.6 Garbage as a Data Source 46
2.2.7 Boxplot as an Input for Decision Making 46
2.2.8 A Series of Inputs for Strategic Decision Making 48
2.3 Data Transformation 48
2.3.1 To Generate Categorical Variables 49
2.3.2 To Generate Dummy Variables 51
2.4 Biserial Correlation Analysis 51
2.5 One-Way Tabulation of a Variable 53
2.6 Two-Way Tabulations 54
2.6.1 Measure of Associations for Bivariate Categorical Variables 58
2.6.2 Other Measures of Association Based on a $2 \times 2$ Table 58
2.6.3 Measures of Association Based on a $I \times 2$ Table 64
2.7 Three-Way Tabulation 67
2.7.1 Conditional Measures of Association for a $2 \times 2 \times 2$ Table 69
2.7.2 Conditional Odds Ratio for an $I \times J \times 2$ Table 70
2.8 Special Notes and Comments 74
2.9 Special Cases of the *N*-Way Incomplete Tables 77
2.10 Partial Associations 80
2.11 Multiple Causal Associations Based on Categorical Variables 81
2.11.1 Theoretical and Empirical Concepts of Causal Associations 81
2.11.2 Multidimensional Frequency Table 85
2.12 Seemingly Causal Model Based on Categorical Variables 89
2.12.1 Causal Association Based on (X1, X2, Y1) or (X1, Y1, Y2) 90
2.12.2 Causal Association Based on (X1, X2, Y1, Y2) 91
2.12.3 Causal Association Based on Multidimensional Variables 94
2.13 Alternative Descriptive Statistical Summaries 95
2.13.1 Application of the Object "Descriptive Statistics and Test" 95
2.13.2 Application of the Object "Graph. . ." 102
2.14 How to Present Descriptive Statistical Summary? 107
2.14.1 DSS Based on a Set of Zero–One Indicators 107
2.14.2 Two-Dimensional DSS of Proportions 108
2.14.3 Multidimensional DSS of Proportions 108
2.14.4 DSS Based on a Set of Agung–Likert Scale Attributes 109
2.14.5 DSS Based on a Set of Numerical Problem Indicators 110
2.14.6 Additional Descriptive Statistical Summaries 111

2.15 General Seemingly Causal Model 111
2.16 Empirical Studies Presenting Descriptive Statistical Summaries 112
2.16.1 Studies in the Field of Nutrition 112
2.16.2 Studies in Public Health 114
2.16.3 Selected Experimental Studies 114
2.16.4 Studies in Public Relations 114
2.16.5 Studies on Other Population Problems 115

**3 One-Way Proportion Models 117**
3.1 Introduction 117
3.2 One-Way Proportion Models Based on a 2 × 2 Table 117
3.2.1 Regression Functions 118
3.2.2 Binary Logit Functions 119
3.2.3 Odds Ratio Statistics 120
3.3 Binary Choice Models Based on a $K \times 2$ Table 121
3.3.1 Binary Logit Models 121
3.3.2 Binary Multiple Regressions 122
3.4 Binary Logit Models Based on *N*-Way Tabulation 122
3.4.1 Binary Logit Models Based on Three-Way Tabulation 122
3.4.2 Binary Choice Models Based on Higher Dimensional Tables 124
3.5 General Binary Choice Models 124
3.5.1 Binary Multiple Regression Model 125
3.5.2 The Wald Test 127
3.5.3 Binary Logit Models 134
3.5.4 Binary Probit Models 144
3.5.5 Binary Extreme-Value Models 147
3.6 Special Notes and Comments 151
3.6.1 The True Population Binary Choice Model 151
3.6.2 The Sampled Binary Choice Function 151
3.6.3 Alternative Equation Estimations 152
3.7 Association between Categorical Variables 152
3.7.1 Generating the Dummy Variables 153
3.7.2 Generating a Cell Factor 154
3.8 One-Way Binary Choice Models Based on *N*-Way Tabulation 156
3.8.1 *N*-Way Tabulation without an Empty Cell 156
3.8.2 *N*-Way Tabulation with Empty Cells 157
3.8.3 Testing Hypotheses 157
3.9 Special Notes and Comments on Binary Choice Models 160

**4 *N*-Way Cell-Proportion Models 165**
4.1 Introduction 165
4.2 The *N*-Way Tabulation of Proportions 165
4.2.1 A 2 × 2 Table of Proportions 165
4.2.2 A I × J Table of Proportions 167
4.3 The 2 × 2 Factorial Model of Proportions 168
4.3.1 Pure Interaction Models 168
4.3.2 Interaction Models with a Main Factor 170

4.3.3 Interaction Models with Both Main Factors 174
4.3.4 Additive Binary Choice Models 175
4.4 $I \times J$ Factorial Models of Proportions 176
4.4.1 Interaction Models 176
4.4.2 Special Notes and Comments 178
4.5 Multifactorial Cell-Proportion Model 180
4.6 Presenting the Statistical Summary 188

**5 *N*-Way Cell-Mean Models 193**
5.1 Introduction 193
5.2 One-Way Multivariate Cell-Mean Models 195
5.2.1 An MCMM without an Intercept 195
5.2.2 An MCMM with Intercepts 195
5.3 *N*-Way Multivariate Cell-Mean Models 197
5.3.1 Two-Way Multivariate Cell-Mean Models 197
5.3.2 Three-Way Multivariate Cell-Mean Model 201
5.3.3 *N*-Way Multivariate Cell-Mean Model 202
5.4 Equality Test by Classification 202
5.5 Testing Weighted Means Differences 208
5.6 Descriptive Statistical Summary 212

**6 Multinomial Choice Models with Categorical Exogenous Variables 213**
6.1 Introduction 213
6.2 Multinomial Choice Models 213
6.2.1 Multinomial Logit Model as a Set of (M − 1) Binary Logit Models 213
6.2.2 Multinomial Logit Model as a Set of M Binary Choice Models 224
6.3 Ordered Choice Models 225
6.3.1 Simple Ordered Choice Models 225
6.4 Concordance–Discordance Measure of Association 231
6.5 Multifactorial Ordered Choice Models 234
6.6 Multilevel Choice Models 241
6.6.1 Two-Level Choice Models 241
6.6.2 Three-Level Choice Models 250
6.7 Special Notes on the Multinomial Logit Model 253
6.8 Selected Population Studies Using Multinomial Choice Models 256
6.8.1 Multinomial Problem Indicators and Gender Equity Indexes 256
6.8.2 Multinomial Problem and Poverty Indicators 259

**7 General Choice Models 263**
7.1 Introduction 263
7.2 Binary Choice Models with a Numerical Variable 263
7.2.1 The Simplest Binary Choice Model 263
7.2.2 Alternative Simple Binary Choice Models 269
7.2.3 Special Notes and Comments 276
7.3 Heterogeneous Binary Choice Models 276
7.3.1 The Simplest Heterogeneous Binary Choice Model 276
7.3.2 General Heterogeneous Binary Choice Model 282

7.4 Homogeneous Binary Choice Models 284
7.4.1 Binary Choice ANCOVA Model with a Numerical Variable 284
7.4.2 Graphical Representation of an ANCOVA Model 287
7.5 General Binary Choice Models 288
7.5.1 Hierarchical Binary Logit Model 288
7.5.2 Nonhierarchical Binary Logit Model 289
7.5.3 Additive Binary Logit Model 290
7.5.4 GBCM with Two Numerical and a Dichotomous Independent Variable 293
7.5.5 GBCM with Two Numerical and a Set of Categorical Independent Variables 297
7.6 Advanced Binary Choice Models 298
7.6.1 Binary Choice Heterogeneous Regressions 298
7.6.2 Binary Choice ANCOVA Model 304
7.6.3 Descriptive Statistical Summaries 307
7.7 Multidimensional Binary Choice Translog Linear Model 307
7.8 Piecewise Binary Choice Models 309
7.9 Ordered Choice Models with Numerical Independent Variables 313
7.10 Studies Using General Choice Models 325
7.11 Two-Stage Binary Choice Model 326

**8 Experimental Data Analysis 329**
8.1 Introduction 329
8.2 Analysis Based on Cell-Mean Models 329
8.2.1 The Simplest Statistical Analysis 330
8.2.2 Special Remarks 331
8.2.3 Application of Multivariate Cell-Mean Models 332
8.3 Bivariate Correlation Analysis 333
8.4 Effects of the Experimental Factors 334
8.5 Effects of the Experimental Factors and Covariates 335
8.5.1 Effects of the Experimental Factors and a Covariate 336
8.5.2 Effects of the Experimental Factors and Two Covariates 342
8.5.3 The Application of Translog Linear Models 346
8.6 Application of the Ordered Choice Models 356
8.7 Application of Seemingly Causal Models 360
8.7.1 The Simplest Seemingly Causal Model 361
8.7.2 Four Pairs of Causal Relationships 363
8.7.3 Five Pairs of Causal Relationships 364
8.7.4 All Pairs Have Causal Relationships 365
8.7.5 Alternative Seemingly Causal Models 368
8.7.6 Special Notes and Comments 369
8.8 Multivariate Analysis of Covariance 369
8.9 Tests for Equality of Medians 372
8.10 The Simplest Experimental Design 376

**9 Seemingly Causal Models Based on Numerical Variables 381**
9.1 Introduction 381
9.2 The Simplest Seemingly Causal Model 382
9.2.1 Bivariate Correlation and the Simplest Linear Regression 382
9.2.2 Scatter Graph with Regression Line 385
9.2.3 Residual Analysis 389
9.2.4 Special Notes and Comments 390
9.3 General Linear Models Based on Bivariate (*X*, *Y*) 391
9.3.1 Continuous Regression Models 391
9.3.2 Discontinuous Regressions 402
9.3.3 Regressions by a Classification Factor 405
9.4 Models Based on Numerical Trivariate 407
9.4.1 Continuous Regressions 407
9.4.2 Regressions by Classification Factors 416
9.5 Regression Analysis Using the Principal Components 417
9.6 Seemingly Causal Models Based on (X1, X2, Y1, Y2) 420
9.7 Seemingly Causal Models Based on (X1, X2, X3, Y1, Y2) 422
9.7.1 The Model with the Dependent Variable Y1 423
9.7.2 The Model with the Dependent Variable Y2 424
9.7.3 The Model with the Dependent Variable X1 424
9.7.4 The Model with the Dependent Variable X3 424
9.8 New Types of Interaction Model 426
9.8.1 Polynomial Interaction Model 426
9.8.2 General Polynomial Interaction Model 428
9.8.3 System Polynomial Interaction Model 430
9.9 Special Cases 431
9.9.1 Predicted Variables and Predictors 431
9.9.2 The Simplest and the Most Complex Seemingly Causal Models 433
9.10 Special Notes and Comments 434
Appendix A.9.1 Hypothetical Data Set 435

**10 Factor Analysis and Latent Variables Models 439**
10.1 Introduction 439
10.2 The Basic Concept of Factor Analysis 440
10.3 The First-Level Latent Variables 441
10.3.1 Generating Additive Latent Variables 441
10.3.2 Interaction Latent Variables 447
10.3.3 Special Notes and Comments 449
10.4 Illustrations Based on Hamsal's (2006) Data Set 450
10.4.1 Generating Latent Variables 450
10.4.2 Latent Variable Regression Models 452
10.4.3 Alternative Latent Variable Models 453
10.5 Selected Cases Based on Ary Suta's (2006) Data Set 458
10.5.1 Multilevel Latent Variables 458
10.5.2 Problems with the Sample Sizes 459

10.6 Evaluation Analysis Based on Latent Variables 462
10.6.1 Ordinal Classification Based on a Latent Variable 462
10.6.2 Composite Index Based on a Latent Variable 463
10.6.3 *N*-Way Tabulation Based on Latent Variables 464

**11 Application of the Stepwise Selection Methods 467**
11.1 Introduction 467
11.2 The Options for the Stepwise Selection Methods 467
11.3 Selection Method for the Numerical Variable Regression Models 469
11.3.1 Two-Way Interaction Stepwise Regressions 469
11.3.2 Three-Way Interaction Stepwise Regressions 473
11.3.3 Application of Multistage Stepwise Selections 476
11.3.4 Alternative Selection Methods 478
11.4 Multifactorial Stepwise Regression Models 480
11.4.1 Multifactorial Cell-Mean Models 480
11.4.2 Multifactorial Heterogeneous Regressions 482
11.4.3 Stepwise ANCOVA Models 491
11.5 Illustrative Stepwise Regressions Based on Mlogit.wf1 495
11.5.1 Classical ANCOVA Models 495
11.5.2 Interaction ANCOVA Models 497
11.6 Special Notes and Comments 503

**12 Censored Multiple Regression Models 505**
12.1 Introduction 505
12.2 Tobit Models 505
12.2.1 Tobit Cell-Mean Models 506
12.2.2 Tobit Regression Models with Numerical Independent Variables 511
12.2.3 Selected Studies Using Tobit Regressions 517
12.3 General Tobit Model 517
12.4 Zero–One Indicator of Censoring 521
12.5 Illustrative Cases of Censored Observations 526
12.5.1 Outliers are Considered as Censored Observations 526
12.5.2 Both Tales of Observations are Considered as Censored Observations 527
12.5.3 Waiting Time and Switching Status Variables 528
12.6 Series for a Censoring Variable 528
12.7 Switching Censored Regressions 531
12.8 Special Notes and Comments 542
Appendix A.12.1 Hypothetical Censored Data, Modified from Faad's (2008) Data Set 543

**References 545**

**Index 551**

# Preface

It is well known that EViews is excellent software for conducting time-series and panel data analyses. However, it has never been considered for doing cross-section data analysis. Based on my own experiences in writing several Indonesian books and papers on data analysis using SPSS, and doing a lot of experiments using EViews, I have found that EViews provides better programs or options for several statistical analysis methods than SPSS does.

The descriptive statistical methods that are very important to mention are specifically the option *Equality Tests by Classification*, which can easily be used to construct various descriptive statistical summaries, by using or inserting any sets of categorical and numerical variables. For inferential statistical methods, EViews provides the Wald test (which can easily be used to test various hypotheses using the model parameters), an object *System* (which can be used to represent a general linear model (GLM, either univariate or multivariate), a structural equation model (SEM), and a seemingly causal model (SCM) – refer to Agung (2009a)), several estimation settings to conduct analysis based on instrumental variables, and STEPLS (stepwise least squares), which has a unique method, namely the combinatorial selection method. Furthermore, EViews also provides many functions, so that anyone can easily generate new series or variables, such as the simplest function @Meansby(arg1, arg2 [ ,s]) for generating the mean of ARG1 by the categorical variable ARG2, and many advanced functions which are beyond the scope of this book.

Furthermore, I have found that all types of model and method for cross-section and experimental data presented in this book can easily be applied to panel data having a large number of observed individuals or objects, by taking into account an additional time *t*-variable. Take note that one of the categorical variables of any of the models presented in this book could be replaced by the categorical time *t* variable. If the panel data have a sufficient number of time-point observations, then the time *t* can be used as a numerical independent variable. Finally, with regard to the piecewise regression models, the models for panel data could have the numerical time *t* variable together with its defined dummy variables. Thus, this book would be an excellent complement or guide for doing analysis based on panel data having a large number of observed individuals or objects.

Each chapter in this book demonstrates the simplest possible data analysis of those presented in the whole chapter. I am very confident that the simplest data analysis (such as the one- and two-way tabulations of the proportions, means, and median), the simplest linear regression of a bivariate numerical variable, and the simplest binary choice model having a single categorical independent variable in particular, as well as simple graphs, can easily be understood by all

readers, especially the statistical users. Furthermore, special notes and comments are also presented for any unexpected or uncommon results. Hence, I would say that this book should be a good guide for undergraduate and graduate students in doing data analysis. And regarding this point, it is important to mention that I have advised many undergraduate students, including my children and grandsons, since 1960.

On the other hand, based on my observations, even graduate students and less-experienced researchers do not think that descriptive statistical analysis is the most useful analysis in an evaluation study. Refer to Chapter 2, specifically the illustrative empirical studies presented in Section 2.15.

It is recognized that all statistical models having numerical dependent variables and their estimation methods based on cross-section data can easily be derived from my first book (Agung, 2009a); thus, the application of those models will not be presented in detail in this book.

Furthermore, note that all models having numerical dependent variables based on any cross-section data, such as the additive and two- and three-way interaction SCMs, could easily be derived from all time-series models presented in Agung (2009a) by using two alternative methods or modifications. And similarly for the instrumental variable models (Chapter 7), nonlinear models (Chapter 10), and the nonparametric estimation methods (Chapter 11).

The first method is to delete the time $t$ variable as well as the lags of endogenous and exogenous variables from the time-series models, and then simpler models would be obtained containing fewer variables. The second method is to replace the time $t$ variable as well as the lags of endogenous and exogenous variables by a relevant set of numerical or dummy variables. Then, under the assumption that the corresponding path diagram is acceptable, in a theoretical sense, various cross-section models, either additive or two- and three-way interaction models, could easily be defined. However, in some cases the path diagrams should be modified to anticipate the structural relationship of the new variables. Many alternative path diagrams are presented in Agung (2009a), and additional illustrative path diagrams will be presented in this book. Finally, all methods for testing hypotheses, specifically the additional or advanced testing hypotheses presented in Agung (2009a: Chapter 9), can be applied directly.

For these reasons, the first book should be considered as a very important complement of this book. Hence, this book will mainly present illustrative data analyses based on models having categorical dependent variables, such as the binary and multinomial choice models, including the ordinal choice models, having categorical or numerical independent variables, as well as both types of independent variable. In addition, selected models having numerical dependent variables, such as the latent variable models and censored dependent variable models, are also presented.

This book contains 12 chapters.

Chapter 1 presents special notes and comments on selected theoretical concepts of statistics, which are misinterpreted by all students and less experienced researchers based on my own observations. Some of the theoretical concepts are the sample space, representative sample, statistic as a function or parameter estimator, hypothesis testing, and random variable and its theoretical normal distribution. In addition, notes on specific groups of variables and causal relationships between variables, in a theoretical sense, are presented, which should be considered even before doing data analysis. Finally, special notes and comments on the

reliability and validity of instruments or tests, as well as on forecasting and the misinterpretation of some sampled statistics, are presented.

Chapter 2 presents simple descriptive statistical summaries which are considered as very important quantitative inputs for decision makers. The statistical summaries presented are frequency tables of the problem (endogenous) indicators by their relevant cause (exogenous) factors, since it is recognized that the causal relationship between variables can be studied using their tabulation based on the transformed categorical variables. In addition, unconditional and conditional measures of association based on *N*-way tabulation are presented. Since, with regard to *N*-way tabulation, EViews also provides the test statistics directly (namely chi-square and likelihood ratio statistics), then this chapter also presents the testing of hypotheses on the multiple associations, as well as on the conditional associations, between categorical variables. Furthermore, this chapter also presents special notes on the statistics-based frequency table with empty cells and incomplete tables. In addition, by defining a specific pattern of causal association between a set of categorical variables, in a theoretical sense, this chapter proposes that the *N*-way tabulation procedure can also be used to test their causal associations. Finally, this chapter demonstrates how to present a descriptive statistical summary based on any set of variables, with examples of empirical studies in various selected fields.

Chapter 3 presents various linear models which have a zero–one indicator as an endogenous variable and a set of dummy variables as exogenous variables generated by either a single or multiple categorical factors, called one-way proportion models. The alternative one-way proportion models presented are the regressions and the binary choice models, starting with the simplest models based on a $2 \times 2$ table. For the model having multiple categorical exogenous variables, it is proposed to generate a cell factor, namely CF, so that any multifactorial models can easily be presented as one-way proportion models. In addition, this chapter demonstrates that a lot of regressions and logistic functions can easily be written based on a cell-proportion tabulation, which are called subjective-regression and subjective-logistic functions. Special notes and comments on the three alternative binary choice models, namely the binary logit, probit, and extreme-value models, are presented. Finally, this chapter presents a one-way proportion model based on the *N*-way tabulation with empty cells and incomplete frequency tables.

Chapter 4 presents various two-way cell-proportion models or bi-factorial design models of a zero–one endogenous variable, in the form of multiple regressions of binary choice models, with categorical exogenous variables. The main objectives of these models are to test hypotheses on the main and interaction effects of categorical exogenous variables on zero–one endogenous variables. For illustration purposes, three types of nonhierarchical model, a full-factorial or hierarchical model, and an additive model are presented. To generalize, multifactorial design models are presented; however, they are treated as if they are bi-factorial design models. For example, based on factors $A$ and $B$, three alternative nonhierarchical models with designs $[A^*B]$, $[A + A^*B]$, and $[B + A^*B]$ and one hierarchical model with a design $[A + B + A^*B]$ can be presented. For the multifactorial designs, the factors $A$ or $B$ can represent two cell factors, namely CF1 or CF2, which are generated based on the subsets of the multifactors. Finally, this chapter presents special multifactorial binary choice models with unexpected statistical results, which are related to the special incomplete frequency tables.

Chapter 5 mainly presents *N*-way or multifactorial multivariate cell-mean models; however, for data analysis, the models would be considered or presented as one-, two-, or three-way

multivariate cell-mean models. In addition, the test for equality of means, variances, and medians of a single numerical variable by various types of categorical factor are presented.

Chapter 6 presents the application of two types of multinomial choice model. The first type of model is a set of one-way and multifactorial binary choice models, and the second type are the one-way and multifactorial ordered choice models. In addition, simple two- and three-level choice models are presented. Finally, special notes on the true multinomial logistic model are presented.

Chapter 7 presents the application of various binary and ordered choice models having numerical exogenous variables, or both numerical and categorical variables, called general choice models (GCMs), starting with various simple GCMs with a single numerical exogenous variable. The models can then easily be extended to a lot of choice or discrete models, both linear and nonlinear GCMs. Special notes and comments on the acceptability of estimates, in a statistical sense, are presented supported by residual analysis.

Chapter 8 demonstrates the applications of various statistical models based on an experimental data set by using the original measured variables and their transformed variables, such as the natural logarithm of the variables, and the zero–one indicators as well as ordinal variables (even though the data only have numerical variables), in addition to treatment factors. In fact, the models, such as the binary and ordered choice models, have been presented in previous chapters. Furthermore, the application of various causal models is also presented, such as MANOVA, homogeneous regression (MANCOVA), heterogeneous regressions, and the system equations, based on numerical variables.

Chapter 9 presents SCMs based mainly on numerical variables, which are most likely do not have pure causal relationships, since the data used is cross-section data. The SCMs presented start with the simplest linear regression in a two-dimensional space, namely SLR_2, based on a pair of numerical variables, and the patterns of their possible relationships are graphically presented in a two-dimensional coordinate system. This is then extended to the simplest linear regression in three-dimensional space, namely SLR_3, and SLR in the $k$-dimensional space, namely SLR_$k$, for $k > 3$, called a *hyperplane* in an abstract space. Finally, various SCMs are presented that correspond to specific path diagrams which represent the causal relationships defined theoretically based on a set of numerical variables.

Chapter 10 presents illustrative examples on how to develop a latent variable based on a set of defined measurable or observable variables or attributes, by using the object *Factor* or *Factor Analysis*, which is available only in EViews 6, and the principal component method for previous versions of EViews. However, this chapter will not present the application of latent variable models in detail, since they can be considered exactly the same as all models based on numerical variables, which are presented in previous chapters and in Agung (2009a). Hence, this chapter only presents single-stage and multistage factor analysis in generating latent variables, selected latent variable models with special notes and comments, and the evaluation or policy analysis based on latent variables.

Chapter 11 demonstrates various acceptable and unexpected regressions obtained based on the same set of three search regressors, in particular based on a larger number of search regressors, by using a multistage stepwise selection method. Since it is well known that the effect of an exogenous (cause, source, upstream, or independent) variable on an endogenous (downstream, impact, or dependent) variable, in general, is theoretically dependent on other exogenous variables, then a two-way interaction model should be applied, and selected or limited three-way interactions may be used as additional independent variables if and only if

the corresponding three main factors are completely correlated or associated, in a theoretical sense. As an extension of all the regressions presented, various similar regressions can easily be developed using transformed variables, such as the semilog and translog linear or nonlinear models, as well as the bounded regression models having $\log[(Y-L)/(U-Y)]$ as an dependent variable, where $L$ and $U$ are the lower and upper bounds of any numerical endogenous variable $Y$.

Finally, Chapter 12 presents unexpected empirical findings based on various Tobit and censored regression models, starting from the simplest models, such as the censored mean and cell-mean models, and censored regression models with one and two numerical variables, are introduced by using the fitted or index values variables, either as dependent or independent variables. Finally, it is found that ARCH and GARCH/TARCH models should be applied based on a switching censored regression.

In addition, I present special notes and comments, most of which are not presented in statistical books as well as research methods. Thus, it is expected that readers will have a better picture of the limitations of the various models, as well as the problems found in doing data analysis, specifically in obtaining alternative acceptable or good-fit models based on sampled data that happen to be selected or are available for the researcher. Refer to the notes and comments about a sample presented in Chapter 1. Furthermore, the advice to readers is to read the special notes and comments presented in Agung (2009a).

I wish to express my gratitude to the Graduate School of Management, Faculty of Economics, University of Indonesia, and the Ary Suta Center, Jakarta, for providing a rich intellectual environment and facilities that were indispensable for the writing of this text. In the process of writing the draft of this book using the PC, I would like to thank the staff of the Graduate School of Management, specifically Tridianto Subagio and Asep Saepul Hayat, who gave great help if I had problems with the software.

In the process of writing this applied statistical book in English, I am indebted to my daughter, Ningsih Agung Chandra, and my son, Darma Putera, for their time in correcting my English. My daughter has a Bachelor of Science from the Department of Biostatistics, School of Public Health (BSPH), the University of North Carolina at Chapel Hill, USA, and a Master's Degree in Communication Studies (MSi) from the London School of Public Relations – Jakarta (LSPR). Now, she is a senior lecturer and thesis coordinator of the graduate program, as well as advisor of both undergraduate and graduate programs at LSPR. In addition, she is also the PR & Communication Manager of the Macau Government Tourist Office (MGTO) Representative in Indonesia, and her profile can also be found through Google by typing her complete name – Martiningsih Agung Chandra. My son has an MBA from De La Salle University, Philippines, and a BSc in Management from Adamson University, Philippines, and he had been in the USA for more than 5 years while I was studying in the USA. Now, he is the director of Pure Technology Indonesia, which has a company (Pure Technology Philippines) in Makati, as a subsidiary of Pure Technology Indonesia.

Finally, I would like to thank Dr Esther Levy, the reviewers, editors, and all the staff at John Wiley & Sons (Asia) Pte Ltd for their hard work in getting this book, and my first book, *Time Series Data Analysis Using EViews*, to publication.

# 1

# Misinterpretation of Selected Theoretical Concepts of Statistics

## 1.1 Introduction

It is recognized that most undergraduate and graduate students do not have sufficient knowledge of the basic theoretical concepts of statistics or mathematical statistics in general, such as the concepts of sample space, representative sample, a statistic as a parameter estimator, testing hypothesis, and random variable and its theoretical (normal) distribution. Thus, they think that they have to prove the theoretical concepts of statistics by using only a sample data set, for instance, error terms of statistical models should have independent and identical normal distributions. In fact, some of them even think that the cross-section data of a numerical variable should be tested whether or not it has a normal distribution, before doing further data analysis.

On the other hand, they think that they have to select a representative sample for their theses or dissertations. This does not have a clear meaning, as in fact there is no sampling method or guide on how to select a representative sample. Agung (1992a) states that it is better not to use the term "representative sample" anymore, since it can be misleading. It is well known that researchers will most likely select a nonprobability sample, specifically a convenient sample where each respondent has a convenient time in giving "good" responses. In other words, most sample survey researches do not use pure random samples, since researchers never take into account a complete list of the population. On the other hand, a random sample is basically defined as a sample where each individual in the population has an equal probability of being selected. In fact, even some of my graduate students for their theses and dissertations should be using a special sampling method, called the *friendship sampling method* (Agung, 2008a), since they should interview managers or high-ranking persons, who have very limited time and most likely do not want to participate in the study, or they are using their friends as the research objects.

Furthermore, it is recognized that most books in applied statistics do not present or discuss the sample space with simple and detailed illustrations. Hence, students or readers never clearly know the limitation of a sample data set for estimating the true value of population parameters. On the other hand, several sampled statistics can be misinterpreted, such as a causal relationship between a pair of variables should be proven using a simple regression,

*Cross Section and Experimental Data Analysis Using EViews*, First Edition. I Gusti Ngurah Agung.

the standard error of a variable, a sample size has to be estimated using a statistical formula, the reliability Crownbach $\alpha$, which in fact is a consistency coefficient, and validity of an instrument data collection, which are in fact computed based on a sample of individuals that happen to be selected by the researchers.

For this reason, the following sections present some notes and comments on selected theoretical concepts of statistics, as well as sampled statistical values, which are considered as very important supporting knowledge in giving values to the statistical results based on a sample data set.

## 1.2 What is a Population?

It has been recognized that a population can be thought of as a complete set of individuals, a complete set of characteristics or variables, or as a complete set of scores, values, or measurements of variables. For these reasons, the following alternative definitions of a population are proposed. On the other hand, a hypothetical population will be introduced later, corresponding to any nonrandom samples which have been used in most or almost all sample survey researches.

**Definition 1.1**

A population is defined as a complete set of all individuals having specific characteristics defined by a researcher, such that each individual can be perfectly classified into whether or not the individual is a member of the population.

**Definition 1.2**

A population is defined as a complete set of all possible characteristics or variables of the observed individuals.

**Definition 1.3**

A variable is a characteristic of a set of individuals, which can have different scores/values/measurements for different individuals in the set.

**Definition 1.4**

A population is a complete set of multidimensional quantitative and qualitative scores, values, or measurements of all possible variables, which could give a complete data or information to a researcher. In other words, a population is a complete set of quantitative and qualitative scores/values/measurements of all possible defined variables.

## 1.3 A Sample and Sample Space

### *1.3.1 What is a Sample?*

**Definition 1.5**

A sample is a finite subset of a defined population. According to Definition 1.4, then, the sample data set, which will be called "sample" for short, will be defined as a finite set of quantitative

and qualitative scores, values, or measurements which happen to be selected by or are available for a researcher.

Take note that in any statistical data analyses, researchers or analysts would always consider a very small set of scores/values/measurements for a limited number of all possible variables or indicators. Researchers will never be observing or measuring a whole population, either as the complete sets of individuals, variables or characteristics, as well as scores/values/measurements. For this reason, every reader should be very confident that researchers will never have the total or complete information of a population, in general.

Corresponding to a sample survey, the population could be classified as a *sample population* and a *target population.* A sample population is defined as the population, from which a sample will directly be selected, by using a specific selection method, such as a multistage sampling method and others (see Kish (1975)). The target population is defined as a much larger population than the sample population for which the statistical results are predicted, estimated, or assumed to be applicable, in a statistical sense. Take note that this statement on the target population is an abstract or theoretical statement, which cannot be proven empirically. Refer to the notes presented in the following sections, as well as the notes in Agung (2009a: Section 2.14).

### *1.3.2 What is the Sample Space?*

**Definition 1.6**

*A sample space* is a set of all possible samples having a fixed size, say $n$, which could be selected from a defined population.

Note that a researcher will never observe a sample space; thus, a sample space should be considered as an *abstract theoretical concept*. Corresponding to the population as a set of a complete individuals, then, any samples of size $n$ would generate or have a sample space consisting of all possible sets of $n$ individuals in the population of size $N > n$. In this case, then, *the sample space of n individuals from the population of size N* can be presented as follows:

$$\Omega(n|N) = \{S_i(n);\ i = 1, \ldots, C(n,N)\} \tag{1.1a}$$

with

$$C(n,N) = \frac{N!}{n!(N-n)!} \tag{1.1b}$$

where $S_i(n)$ indicates the $i$th sample of $n$ individuals, $C(n,N)$ indicates a combination of $n$ out of $N$ individuals, and $k!$ (i.e. $k$-factorial) equals $(1 \times 2 \times 3 \times \ldots \times k)$, and it is defined that $0! = 1! = 1$.

In general, a sample space would have such a large number of elements. The following examples present the concepts of sample space and statistic, with simple illustrations.

**Example 1.1** Sample spaces of a population of size $N = 5$

Suppose a sample of size $n$ is to be selected from a population of size $N = 5$ having identification numbers 1, 2, 3, 4, and 5; then, the following alternative sample spaces, namely the *sample space of individuals*, could be obtained.

1. For $n = 1$, the sample space is

$$\Omega(n = 1|N = 5) = \{1, 2, 3, 4, 5\}$$

   which is equal to the population. Note that the sample of size $n = 1$ can only provide 20% of the total information (i.e., one out of five) individuals in the population, which happen to be selected.
2. For $n = 2$, the sample space is

$$\Omega(n = 2|N = 5) = [\{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{2,3\}, \{2,4\}, \{2,5\}, \{3,4\}, \{3,5\}, \{4,5\}]$$

   where $\{i, j\}$ indicates a set of two individuals in the population, which will be written or presented as

$$\Omega(n = 2|N = 5) = \{12, 13, 14, 15, 23, 24, 25, 34, 35, 45\}$$

   Note that this sample space has $C(2, 5) = 10$ members, namely pairs of individuals, or a set of 10 observation pairs, which is larger than the population size $N = 5$. A member of this sample space could provide information only 40% out of the five individuals in the population. Compared with a sample size of $n = 1$, each member of this sample space can provide more information.
3. For $n = 3$, the sample space of size $C(3, 5) = 10$ will be

$$\Omega(n = 3|N = 5) = \{123, 124, 125, 134, 135, 145, 234, 235, 245, 345\}$$

   Similar to the previous sample size of $n = 2$, each member of this sample space can provide more information, compared with each member of the sample space $\Omega(n = 2|N = 5)$.
4. For $n = 4$, the sample space of size $C(4, 5) = 5$ will be

$$\Omega(n = 4|N = 5) = \{1234, 1235, 1245, 1345, 2345\}$$

   Compared to the previous sample spaces, each member of this sample space can provide a lot more information, since only one individual is not counted or observed.
5. For $n = 5$, the sample space of size $C(5, 5) = 1$ will be

$$\Omega(n = 5|N = 5) = \{12345\}$$

   In this case, the sample is in fact the whole population. In practice, however, one will never observe the whole population, especially for a large population size, since it will be very expensive and time consuming. □

**Example 1.2** A sample as a member of the sample space

For illustration purposes, Table 1.1 presents the sizes of sample spaces by selected population and sample sizes. Based on this table, the following notes and comments are presented.

**Table 1.1** Illustration of the size of sample spaces by population and sample sizes

| $N$ | 25 | 25 | 50 | 50 | 100 | 100 | 100 |
|---|---|---|---|---|---|---|---|
| $n$ | 2 | 4 | 2 | 4 | 2 | 4 | 5 |
| $\Omega$ size | 300 | 12 650 | 1225 | 230 300 | 4950 | 3 921 225 | 75 287 520 |

Table 1.1 demonstrates that, by increasing the sample size, the sizes of the sample spaces are increasing by a very large quantity. For example, by taking a sample of size $n = 2$ (or 2%) from a population of size $N = 100$, a sample space of size 4950 should be considered; by taking a sample of size $n = 4$ (4%), a sample space of size 3 921 225 should be considered; and by taking a sample of size 5%, a sample space of size more than 75 million should be considered. □

**Example 1.3** The mean-sample-space

Corresponding to the sample space $\Omega(n = 2|N = 5)$ in Example 1.1, let $Y_i$, $i = 1, 2, 3, 4, 5$ be the scores for the $i$th individual; then, it will be considered a sample space of the means or average values, called the *mean-sample-space*, as follows:

$$\begin{aligned}\Omega_{\mathrm{m}}(n = 2|N = 5) &= \{(Y_1 + Y_2)/2, (Y_1 + Y_3/2, \ldots, (Y_4 + Y_5)/2\} \\ &= \{(Y_i + Y_j)/2 | i < j = 1, 2, 3, 4, 5\}\end{aligned}$$

If a sample were to be selected, then the sampled mean selected will be one out of the 10 elements in the sample space, which, most likely, will not be equal to the population mean or the *mean parameter*. However, it is easy to derive that the average value of these 10 average values would equal to the mean parameter, as follows:

$$\{(Y_1 + Y_2)/2 + (Y_1 + Y_3)/2 + \cdots + (Y_4 + Y_5)/2\}/10 = \sum_{i=1}^{5} Y_i/5$$

This simple illustration can directly be generalized to the sample space of the means $\Omega_{\mathrm{m}}(n|N)$, for the population of size $N$ with a fixed sample size $n$. This general mean-sample-space will be presented as

$$\begin{aligned}\Omega_{\mathrm{m}}(n|N) = \Omega_{\mathrm{m}}(n) &= \left\{ \sum_{i=1}^{n} Y_i/n | N \right\} \\ &= \left\{ \sum_{i=1}^{n} y_i/n | \forall Y_j, j = 1, 2, \ldots, N \right\}\end{aligned} \tag{1.2}$$

A special case of this mean-sample-space is the sample space of the proportions called the *proportion-sample-space*, which is obtained if and only if $Y_j$ has a value of zero or one for all $j = 1, 2, \ldots, N$. This proportion-sample-space can be written in a simple form as

$$\Omega_{\mathrm{p}}(n|Y_j = \{0, 1\}, j = 1, \ldots, N) = \{0, 1/n, 2/n, \ldots, (n-1)/n, 1\} \tag{1.3}$$

where $\Omega_p = 0$ indicates that all individuals in the sample have zero scores, and $\Omega_p = 1$ indicates that all individuals in the sample have scores of ones, under the condition the numbers of zeros ($N_0$) and ones ($N_1$) in the population are greater than $n$, namely $N_0 > n$ and $N_1 > n$.

These sample spaces have a $C(n, N)$ total number of element, as presented in (1.1); however, many or a lot of subsample spaces have identical average observed values. Note that the proportion-sample-space $\Omega_p(n|N)$ has only $(n + 1)$ possible values from a set of $C(n, N)$ possible samples. This condition indicates that the sample space of the means, as well as the sample space of the proportion, has a distribution. □

### *1.3.3 What is a Representative Sample?*

In a sample survey, in general, a very large size of sample space should be considered. Refer to the sample space of the sample of size $n = 5$, which could be selected from a population of size $N = 100$, presented in the previous example. So this is what a sample is, a representative sample in particular. An illustration is provided in the following example.

**Example 1.4** A member of the proportion-sample-space

With regard to the proportion-sample-space in (1.3), a sample of any sample size will provide only a single value of proportion, say $k/n$, for a fixed value of $k$, out of the $(n + 1)$ possible proportions. How could you know (or be sure) that the observed proportion is equal to the proportion in the population?

For illustration purposes, let us refer to the sample space $\Omega(n = 2|N = 5)$ presented in Example 1.3. In this case, the sample of size 40% of the population should be considered as a relatively large sample size. Even though the sample is large, every reader should be confident that the 40% selected individuals cannot be representing the other 60% in all of their characteristics. Then by considering a zero–one population of size $N = 5$, namely {1, 1, 1, 0, 0}, the proportion-sample-space is $\Omega_p(n = 2) = \{0, 1/2, 1\}$, as presented above. Note that the proportion of 1 (ones) in the population is 0.60 (i.e., the population mean or the value of the mean parameter), which does not in the sample space. In fact even, by taking a random sample of size $n = 4$, the proportion-sample-space will be $\Omega_p(n = 4|N = 5) = \{1/2, 3/4\}$, since there are only two zeros and three ones in the population, and this sample space also does not contain the value of 0.60. Hence, it is very clear that a sample cannot be representing the population.

Finally, by considering the sample space $\Omega_p(n = 5|N = 100)$, a sample of size 5% should be selected out of more than 75 million possible samples, as presented in Table 1.1. Whatever the sampling method applied, it should be realized that there will never be a representative sample, especially if a sample of size less than 5% were to be selected from a population of size greater than 100. □

On the other hand, the term "representative sample" is, in fact, not an operational statement or terminology, since there is no method or guide on how to select a representative sample. In other words, the term "representative sample" is an abstract statement, which could be misleading. For this reason, Agung (1992a) suggested no longer using this terminology. Furthermore, it is suggested to use any sampling methods which are considered feasible for the study, such as snowball, purposive, convenience, and friendship (*kekerabatan*, in Agung (2008a)). The friendship sampling method should be applied by my students for their dissertations, since they

have to interview managers or high-ranking persons, who are most likely not willing to participate in a study, particularly in Indonesia.

Even though a random sample is required for the inferential statistical analysis, as well as for the statistical theory in general, it has been recognized that a pure random sampling method can never be applied in practice. Note that the pure random sampling method can be applied only if a complete list of individuals in the population is known. However, in practice, a complete list of the individuals in a population is never taken into account by a researcher, then it is suggested the term "random sampling method" should not be used any more in a research proposal. Every reader should be very confident in using any nonprobability sampling method, such as quota, snowball, and convenient sampling methods. For this reason, I propose the following definitions.

**Definition 1.7**

A sample is an element of a sample space *which happens to be selected* from the corresponding sample space.

**Definition 1.8**

A sample data set is a set of multidimensional scores, values, or measurements of a finite number of variables *which happen to be selected by* or *are available for* a researcher.

Under these definitions, every reader should be very confident that sample data cannot provide complete information on the corresponding population, such as the true values of the basic parameters, as well as the true population models. Refer to the notes and comments presented in the following sections and in Agung (2009a: Section 2.14).

### *1.3.4 Relationship between the Sample Space, Population, and a Sample*

Based on the characteristics of the sample space, population, and a sample previously presented, Figure 1.1 illustrates their relationships. This figure shows that the sample space

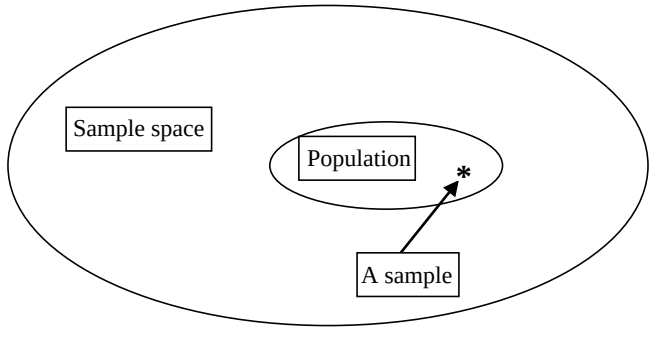



**Figure 1.1** Relationship between of the number of elements in a sample space and its corresponding population and a selected sample.

has a much greater number of elements than the population, and a sample is an *element* (or a *point*) of the sample space which happens to be selected by a researcher. Refer to the illustrative examples presented in Table 1.1, specifically the sample space of size 75 287 520 corresponding to a sample of size on $n = 5$ selected from a population of size $N = 100$. Thence, it is not right to say that a sample is a representative of the corresponding population or, moreover, the sample space.

In empirical statistics, the data analysis is done based on a sample only. On the other hand, the theoretical statistics derive all theorems, statistics, and their distributions based on the sample space, specifically the sample space of a random sample, namely the *random sample space*. See the following sections.

## 1.4 Distribution of a Random Sample Space

In probability or statistical theory, if the sample is a pure random sample, then the proportion-sample-space $\Omega_p$ in (1.3) has the *binomial probability distribution function*:

$$\Pr(\Omega_p = k/n) = C(k,n)p^k q^{n-k} \tag{1.4}$$

where $p$ is the proportion of ones and $q$ is the proportion of zeros in the population, such that $p + q = 1$.

Furthermore, Conover (1980: 544) stated that the binomial probability distribution function can be approximated by a normal distribution for $n > 20$. For illustration purposes, suppose we have the set of values {1, 1, 1, 0, 0} for the population of size $N = 5$ presented above; then, based on the sample space $\Omega_p(n = 2|N = 5)$, the sample space of proportions is $\Omega_p(n = 2|N = 5) = \{0, 1/2, 1\}$, where the number of the zero-proportion is only one $= C(2, 2)$, the proportion of 0.50 is a multiple of $C(1, 3) \times C(1, 2) = 3 \times 2 = 6$, and the proportion of ones is a multiple of $C(2, 3) = 3$. Hence, there are $10 = C(2, 5)$ possible samples of individuals, but three possible proportions. In this case, the sample space has a distribution {0.1, 0.6, 0.3}, which can presented as

$$\Pr(\Omega_p = 0) = 0.1,\ \Pr(\Omega_p = 1/2) = 0.6, \quad \text{and} \quad \Pr(\Omega_p = 1) = 0.3$$

with the following general equation or restriction, based on a random sample of size $n$ from the population of size $N$:

$$\begin{aligned} &\Pr(\Omega_p = k/n) = p_k,\ k = 0, 1, \ldots, n \\ &\sum_{i=0}^{n} \Pr(\Omega_p = k/n) = \sum_{i=0}^{n} p_k = 1,\ p_k \geq 0,\ \forall k \end{aligned} \tag{1.5}$$

Similarly, for the mean-sample-space in (1.2) we will have

$$\sum_i \Pr(\Omega_m = m_i) = 1,\ \infty < m_i < \infty,\ \forall i \tag{1.6}$$

where $m_i$ indicates the $i$th mean or element in the mean-sample-space. In this case, however, the number of distinct mean values cannot be identified.