许迎军 ◎ 著

语言研究统计方法

Statistical Methods in Linguistic Studies

语言研究统计方法

Statistical Methods in Linguistic Studies

许迎军 著

国防工业出版社

·北京·

内 容 简 介

本书介绍了语言研究统计学的基本原理、研究设计和统计方法，内容包括参数估计、假设检验、方差分析、线性回归及相关等，旨在引导学习者领悟语言研究的基本统计思想和方法，学会运用统计数据分析说明语言教学现象。本书统计知识体系完整，统计技能涵盖全面，语言研究应用性强。

本书可作为外语专业本科生、研究生统计学课程的教材，也可供博士和硕士研究生、外语教师、语言研究工作者以及教育科研人员进行定量研究设计和数据分析时阅读参考。

图书在版编目(CIP)数据

语言研究统计方法/许迎军著. —北京：国防工业出版社，2012.1

ISBN 978-7-118-07842-8

Ⅰ.①语… Ⅱ.①许… Ⅲ.①语言统计—统计方法 Ⅳ.①H0—05

中国版本图书馆 CIP 数据核字(2011)第 274734 号

※

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号 邮政编码 100048)

北京嘉恒彩色印刷有限责任公司

新华书店经售

*

开本 880×1230 1/32 **印张** 6½ **字数** 187 千字

2012 年 1 月第 1 版第 1 次印刷 **印数** 1—3000 册 **定价** 35.00 元

Preface

Statistical Methods in Linguistic Studies is for teachers, graduates and postgraduates who major in applied linguistics, instruction and curriculum or other disciplines in education even without a strong background of mathematics aimed at helping students understanding basic concepts of research design and statistics. The book is an introductory text illustrating basic statistical theories and examining statistical tests with a variety of situations where the application of statistical methods would be appropriate and the statistical procedures are easy to follow. Many examples and exercises will assist learners to realize some of the appropriate statistical models and the potential uses of statistics and how statistics can be important to students in their present thesis writing or future employment.

Finally, I would like to thank my colleagues, friends and students who gave me helpful suggestions and encouragement.

Yingjun Xu
October, 2011

Contents

CHAPTER 1 INTRODUCTION

Nowadays the study of statistics has become more popular than ever before. The increasing availability of computers and statistical software packages has enlarged the role of statistics as a tool for empirical research. As a result, calculations and quantifications have pervaded all professions. This is the age of facts, figures and statistics. Today, college students in almost all disciplines are required to take at least one statistics course. As Wells① observed that statistical thinking will be one day as necessary for efficient citizenship as the ability to read and write.

1.1 Overview of Statistics

This introductory chapter explains the basic terms and concepts of statistics, which will bridge our understanding of the concepts and techniques presented in subsequent chapters.

1.1.1 Needs of statistics in linguistic studies

Anyone who listens to the radio, watches television, and reads books, newspapers and magazines cannot help but be aware of statistics. As you begin this course you may well wonder what lies in store for you. Why does linguistics need statistics? As is known, the primary data-source judgments about the well-formedness of sentences

① Herbert George Wells (1866—1946) was a celebrated British thinker and writer, author of novels and science fiction.

usually come from linguists themselves who make either-or decisions. The data simply do not call for or lend themselves to, the assignment of numerical values which need to be summarized or from which inferences may be drawn. Generative grammar, however, despite its great contribution to linguistic knowledge over the past 25 years, is not the sole topic of linguistic study. There are other areas of the subject where the observed data positively demand statistical treatment. (Woods, et al. ,1986). In linguistic studies, as in other sciences such as psychology, sociology and natural sciences, statistical techniques are concerned with linguists, teachers and students who need to appreciate and evaluate literature on statistical analysis and who will be interested in their own research that requires the collection of quantitative data that demands analysis and statistical treatment.

1.1.2 Definition of statistics

The term *statistics* is used in two different ways. As a plural noun, *statistics* refers to numbers or numerical facts. The numbers that represent the scores of students in proficiency examination, frequencies of disorders in a sample of 560 language-impaired individuals, the percentage of pass rate in the test of English majors TEM8 of the university are examples of statistics in this sense of the word. As a singular noun, *statistics* refers to the methods used to analyze numerical data, and to draw conclusions from them. You can say that "statistics *is* the art and science of analyzing numerical data." In this sense of the word, statistics is defined as follows.

Statistics is the science of data analysis; that is, statistics is concerned with scientific methods for collecting, organizing, presenting and analyzing sample data as well as drawing valid conclusions about population characteristics and making reasonable decisions on the basis of such analysis (Iman & Conover, 1985).

1.1.3 Types of statistics

Broadly speaking, statistics can be divided into two kinds: descriptive statistics and inferential statistics.

Descriptive statistics involves methods for tabulating, organizing, displaying, and describing collections of data by tables, graphs and summary measures. These dada may be either *quantitative* variable, such as measures of height or the number of interruptions in a given amount of conversation — variables that are characterized by an underlying continuum — or data may represent *qualitative* variables, such as sex, native language, college major, or aesthetic appeal of an accent, politeness that you can count the *frequencies* or rank in term of *more* and *less*. Large masses of data generally must undergo a process of summarization or reduction before they are comprehensible. Thus, descriptive statistics serves as a tool to describe or summarize or reduce to manageable form the properties of an otherwise unwieldy mass of data.

Inferential statistics consists of a formalized body of methods that use sample results to help make decisions or predictions about a population. These methods are used to solve another class of problems which involves attempts to infer the properties of a large collection of data from inspection of a sample of observations. Thus the aim of inferential statistics is to predict or estimate characteristics of a population from a knowledge of the characteristics of only a sample of the population. The descriptive characteristics of a sample can be generalized to the entire population, with a known margin of error, using the techniques of inferential statistics. For example, we may want to know whether the district's CET average score is significantly higher than the national mean. We may select 1, 000 typical students from different colleges and universities in this district, find

their mean score and make a decision based on this information. The area of statistics that deals with such decision-making procedures is referred to as inferential statistics. Statistics methodology can be illustrated in Figure 1.1 and the relationship between the descriptive statistics and inferential statistics in Figure 1.2.

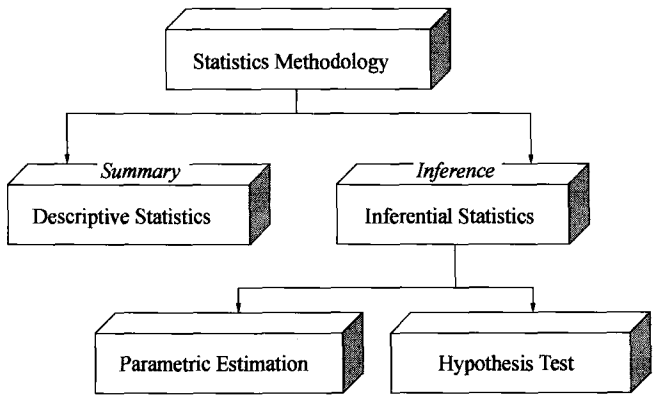

Figure 1.1 Diagram of statistics methodology

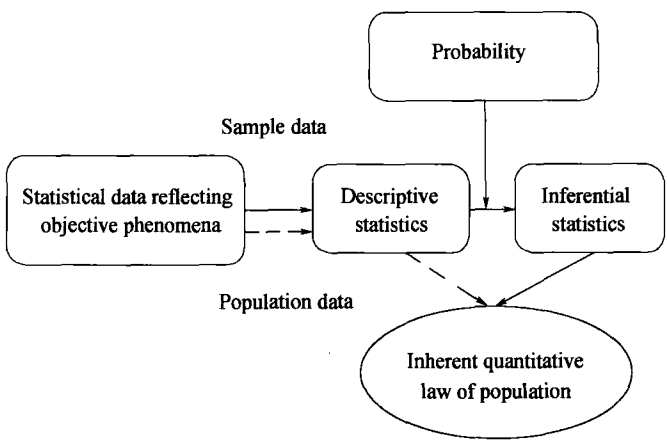

Figure 1.2 Relationship between descriptive and inferential statistics

1.2 Basic Concepts of Statistics

In statistical work, much talk is about populations and samples. It is crucial to understand the meanings of the two terms *population* and *sample* and their difference.

1.2.1 Population and sample

Population refers to the group of observations or all elements including individuals, items, or objects whose characteristics are being studied and about which the investigator wishes to draw conclusions. The population is being studied is also called *target population*.

A *sample* consists of a part of that population. Thus a sample is studied in the hope that it will lead to conclusions about the larger target population.

There are different techniques for sampling. A sample may be random or nonrandom. In a random sample, each element of the population has a chance of being included in the sample. If the chance of being selected is the same for each element of the population, it is called a *simple random sample* or *probability sampling*. However, in a nonrandom sample, some members of the population may not have the chance of being selected in the sample.

One way to select a simple random sample is by a lottery or drawing. For example, if we need to select 5 students from a class of 50, we write each of the 50 names on a separated piece of paper. Then, we place all 50 names in a hat and mix them thoroughly. Next, we draw one name randomly from the hat. We repeat this experiment four more times. The five drawn names make up a simple random sample.

The second procedure to select a simple random sample is to use a table of random numbers (See Appendix A). It is very convenient

to use the table of random numbers. We can start any number in the table, followed by the number down, or designate part of the numbers. For example, 10 students should be selected randomly from the population of 80. To select a simple random sample, we arrange the names of all 100 persons in alphabetic order and compile a two-digit number, from 01 to 80, to each person. Suppose the number 11106 is selected in Line 3, Row 4, we number down along the horizontal direction one by one, and remember this number and its subsequent nine digits—11106, 19620, 67893, 67581, 34534, 09891, 35146, 56269, 94621, 10909. But it should be noted that if the first two-digit number is bigger than 80 or smaller than 01, we can move to the next number. In the above ten numbers, the first two-digit number in 94621 is bigger than 80. Then the next number is 12498. So all the 10 numbers represent 10 subjects selected randomly from the population of 80. Of course, if we have access to a computer, we can use a statistical software package such as MINITAB, to construct a table of more random numbers and select a simple random sample.

If the size of the population is large, the simple random sampling seems tedious and time-consuming. We can use *systematic random sampling* which is more simple and typical. First every member in the population is arranged alphabetically, or in numeric order, or based on some other characteristics. According to the sample size n to be taken, the ratio of population to sample size is $k = N / n$. The starting point a for sampling can be determined randomly. Note that a must be bigger than or equal to 1 ($a \geqslant 1$), and smaller than or equal to k ($a \leqslant k$). Thus, the number sequence for sampling is also determined. The sample n consists of a, $a + k$, $a + 2k$, $\cdots$, $a + (n-1)k$. In systematic random sampling, we first randomly select one member from the first k units. Then every kth member, starting with the first selected member, is included in the sample. For example, we need to select 30 students from a list of 3,000. Since the sample

size should equal 30, the ratio of population to sample size is 3,000/30 = 100. We determine the starting point a, say 220th. Then we select 220, 320, 420, 520, 620, ···, 3120. These randomly selected 30 numbers, which are distributed in the population evenly, constitutes the sample.

Suppose we need to select a sample from the population of a university and we want students with different English proficiency to be proportionately represented in the sample. In this case, instead of selecting a simple random sample or a systematic random sample, we may prefer to apply a different technique *stratified random sampling*. In a stratified random sample, we first divide the population into subpopulations, which are called strata. Then one sample is selected from each of these strata. The collection of all samples from all strata gives the stratified random sample. For example, we first divide the whole population into different groups based on their English levels. We may form three groups of low-, medium-, and high-proficiency. These three subpopulations are usually called strata. Then we select one sample from each subpopulation or stratum. The collection of all three samples selected from three strata give the required sample. It should be noted that the size of the sample selected from different strata are proportionate to the sizes of the subpopulations in these strata and the elements of each stratum are identical with regard to the possession of a characteristic.

Suppose a survey of students' living expense is to be conducted in the boarding school. First, we divide the whole school into n groups, or take each class as a group, or cluster. Make sure that all clusters are similar and, hence, representative of the population. And then certain groups or classes are selected randomly from n groups. Next, certain students from each of these selected groups or classes are chosen randomly. Finally, these selected students are taken as a sample to conduct a survey. This is called *cluster sampling*. Note that all

clusters must be representative of the population.

It is important to note that a random sample is not necessarily an identical representation of the population. Characteristics of successive random samples drawn from the same population may differ to some degree, but it is possible to estimate their variation from the population characteristics and from each other. The variation, known as *sampling error*, does not suggest that a mistake has been made in the sampling process. Rather, *sampling error* refers to the chance variations that occur in sampling; with randomization, these variations are predictable and taken into account in data analysis techniques.

1.2.2 Parameter and statistic

Parameter and statistic are the two technical terms connected with population and sample.

Parameter is a value generated from, or applied to, a population. *Statistic* is a value generated from a sample and sometimes called estimates. Thus, statistics are values derived from sample data, whereas parameters are values that are either derived from, or applied to, population data.

In statistics, population parameters are expressed by Greek letters while sample statistics by English letters. μ stands for population average and $\bar{x}$ represents sample mean; σ stands for population standard deviation and s represents sample standard deviation; π stands for population proportion and p represents sample proportion.

1.2.3 Types of variables

A *variable* is something that may vary, or differ (e.g. language proficiency, self-esteem, motivation, sex, nationality, first language background, intelligence, and language ability). A *constant*, in contrast, has only a single score. For example, if every member of a

sample is male, the "gender" category is a constant. Thus, a variable is a characteristic under study that assumes different values for different elements. In contrast to a variable, the value of a constant is fixed.

There are two types of variables. They are quantitative variables and qualitative variables. *Quantitative variables* are those that are scored in such a way that the numbers, or values, indicate some sort of amount. In other words, they can be measured numerically. They are also called continuous variables. Take for example "height" that is a quantitative or continuous variable. Higher scores on this variable indicate a greater amount of height. The data collected on a quantitative variable are called quantitative data. *Qualitative variables* are those for which the assigned values do not indicate more or less of a certain quality. They are also called categorical variables that cannot assume a numerical value but can be classified into two or more non-numeric categories. So the labels represent qualitative differences in location, not quantitative differences. The data collected on such a variable are called qualitative data.

1.3 Types of Data

Statistical data are obtained in the face of uncertainty because in a certain phenomena the occurrences of the event under study cannot be predicted with certainty. Suppose we study the heights of students aged 10 years in a large city. We cannot predict with certainty the height of an individual student; the data of heights of students constitute statistical data. Further we note that there will be variation in heights of students, though they are of the same age. Uncertainty and variability are two characteristics of statistical data. The classifications of statistical data can be illustrated from different angles.