语言统计学

Statistics in Language Researches

周世界 著

大连海事大学出版社
DALIAN MARITIME UNIVERSITY PRESS

语言统计学

Statistics in Language Researches

周世界 著

大连海事大学出版社

图书在版编目（CIP）数据

语言统计学=Statistics in Language Researches / 周世界著. 大连：大连海事大学出版社，2004.6
ISBN 7-5632-1765-7

I. 语… II. 周… III. 统计语言学－研究－英文 IV. H087

中国版本图书馆 CIP 数据核字（2004）第 043057 号

大连海事大学出版社出版

地址：大连市凌海路 1 号 邮编：116026 电话：0411-84728394 传真：84727996

http://www.dmupress.com E-mail: cbs@dmupress.com

大连海事大学印刷厂印装 大连海事大学出版社发行

2004 年 8 月第 1 版 2004 年 8 月第 1 次印刷

幅面尺寸：140 mm×203 mm 印张：6.875

字数：172 千字 印数：1～400 册

责任编辑：高 炯 封面设计：王 艳

定价：14.00 元

本书由

大连海事大学学术著作出版基金资助出版

The published book is sponsored by

The Academic Works Publishing Foundation
of the Dalian Maritime University

Preface to the First Edition

This book is written as an introductory text primarily for those pursuing their master degrees in English language studies but with no knowledge of statistics. It is my hope that it may also be of interest not only to the practising language teachers but also to the general readers for whom language plays a central part in their activities.

The first book cannot survey all the aspects of statistics in language studies that have been or are being carried out by the statisticians or linguists and only a restricted number of topics are handled. My selection has been governed by the need to expound not only what I consider most important but also what has gained general fame and are regarded as classic techniques or procedures in statistics. Thus, the central purpose of the book is to outline and introduce readers involved to elementary ideas of statistics in language studies which have some relevance for their work, based on my own experience in teaching and research. Growth in the application of statistics to linguistic problems has become rapid and diverse with the rapid development of the computer software such as SPSS in recent years. New and illuminating lines of approach are born out of old ones.

The major emphasis in the book is on the applications of the statistics to language studies that the students are expected to accept in the courses of applied linguistics. To keep the explanations within the comprehension of students with a working knowledge of linguistics but not of the statistics, the book is devoted to the

techniques and process instead of proving equations and theorems of statistics and the main computations of the output are mainly computer-assisted rather than by hand.

No book of this kind could possibly be written without the help and encouragement of others. This book owes much to the instructions and discussions over the years with my beloved professors who have kindly offered their comments and criticisms and read parts of the book: Kong Qingyan, Fan Fengxiang, Tan Wancheng, Shan Yong. Complete drafts of early elephantine versions of this book are read and checked out by the Grade-2003 postgraduate students I have advised: Cao Bing, Zhang Yao, Ma Weiyan, Yang Chuan, Mao Yanwen and Lin Lin, all of whom have made many detailed observations and suggestions. My debts in matters statistic cannot be repaid to all the great statisticians, linguists and language researchers, from whom I have borrowed many ideas, interpretations and illustrating graphs either from the web site or from what they had written in books. Acknowledgements relating to references, quotations and examinations material will be found at the appropriate item in the bibliography.

I am also indebted to Sun Jing for typing the first draft of the book, from whom I got the encouragement to write and to continue. My admiration for her knowledge in medical statistics is equaled by my gratitude for the patience and understanding she has given through the whole time during which the book was being written.

Zhou Shijie

2004.5

Contents

Chapter 1 Some Fundamental Concepts in Statistics

The word "statistics" is used in several different senses. In the broadest sense, "statistics", as a field of study, refers to a range of techniques and procedures for collecting, organizing, analyzing, interpreting data, and making decisions based on the data. This is what courses in "statistics" generally cover.

In a second usage, a "statistic" is defined as a numerical quantity calculated from a sample or samples. Such statistics are used to estimate population parameters. As a result, the term "statistics" sometimes refers to any calculated quantities regardless of whether or not they are from a sample. For example, one might ask about a student's statistics on a test and be referring to his or her each score, the average of all the scores, etc.

Although the different meanings of "statistics" have the potential for confusion, a careful consideration of the context in which the word is used should make its intended meaning clear.

Statistics in language researches—the application of statistics into language studies—is, however, a collection of techniques, as well as the process, used for converting raw linguistic data into information to help linguists, language teachers or others involved to understand how statistics applies to linguistics, to be able to read and understand the statistics presented in professional literature, and to be able to calculate and communicate statistical information to others.

1.1 Descriptive and inferential statistics

One important use of statistics is to summarize a collection of linguistic data, and obtain the descriptive statistics of the data in a clear and scientific way. Descriptive statistics are, therefore, used to describe the basic features of the data in any linguistic study to present quantitative descriptions in a manageable form. They provide simple summaries about the samples and the measures of the data of interest to linguists, language researchers or teachers.

For example, assume a teacher of English gives a test measuring the aptitude to all the students in a school. How might these measurements be summarized? Two basic descriptive methods are applicable: numerical and graphical. Using the numerical approach we might compute the statistics, which convey information about the average degree of the command of English and the degree to which the students differ in their aptitude. Using the graphical approach, however, we might create a histogram, a polygon, etc.

Graphical methods are better suited than numerical methods for identifying patterns in the data while numerical approaches are more precise and objective. Since the numerical and graphical approaches compliment each other, it is wise to use both. Numerical analysis, together with simple graphical analysis, forms the basis of virtually every quantitative analysis of data.

Inferential statistics are, however, distinguished from descriptive statistics in that the former are used to draw inferences about a population based on the statistics from a sample or samples. With descriptive statistics we are simply describing what it is or what the data shows. With inferential statistics, however, we are trying to reach conclusions beyond the immediate data alone. We use

inferential statistics to try to infer from the sample data what the population might be. Or, we use inferential statistics to make judgments of the probability whether an observed difference between groups is a dependable one or one that might have happened by chance in a study. Thus, we use descriptive statistics simply to describe what's going on in our data; we use inferential statistics to make inferences from our data to more general conditions.

1.2 Population and sample

A population is a collection or set of data that describes a phenomenon of interest to linguists or linguistic researchers. We may speak of the population of words in Shakespeare's plays, or the population of the present perfect aspect in a corpus. It is with the characteristics of populations, or aggregates of individual entities, that statistics is most fundamentally concerned.

Populations can be divided into finite and infinite populations. The population of nouns in Shakespeare's plays is finite for the number of entities is fixed and countable. The population of the present perfect aspect is potentially infinite: in theory, at least, we could repeat the use of the present perfect aspect for an infinite number of times.

With a finite population, which is not too large, we may be able to investigate the whole population. But if the population is potentially infinite, or if it is finite but very large, we shall have to be content with samples.

A sample is a subset of data selected from a population. As for the validity of the generalization to the population, selecting samples is based on the assumption that the samples are random, that is, every

unit in the population has an equal chance of being represented in the sample.

1.3 Frequency, relative frequency and probability

It quite commonly arises that we wish to classify a group of linguistic phenomena, putting each unit into one of a set of mutually exclusive classes. The data can then be summarized by giving the frequencies with which each class was observed. Thus the frequency is often called observed frequency and such data are often categorical since each element or individual of the group being studied can be classified as belonging to one or a number of different categories.

It, however, may be more revealing to display the proportions of data falling into different classes, and these can be calculated simply by dividing each frequency by the total frequency. The proportions are called relative frequencies.

The process of making an observation or recording a measurement under a given set of conditions is a trial or experiment, and the outcomes of an experiment are called events. Lengthy observations as to the occurrence or non-occurrence of an event in a large number of repeated trials under the same set of conditions show that for a wide class of phenomena, the number of occurrences or non-occurrences of the event is subject to a stable law. Namely, if we denote by M the number of times the event occurs in N independent trials, then it turns out that, for sufficiently large N, the ratio M/N in most of such series of observations, assumes an almost constant value. Since this constant is an objective numerical characteristic of the phenomena, it is natural to call it the statistical probability of the random event under investigation. Thus, the probability of an event

can be approximated by the proportion of times that the event occurs when the experiment is repeated a very large number of times.

1.4 Classifications of variables and levels of measurement

A variable is any measured characteristic or attribute that differs for different subjects in an experiment or observation. For example, if the frequencies of the perfect aspect of verbs are measured, the frequencies would be a variable. When an experiment or an observation is conducted, some variables are manipulated by the researcher and others are measured from the subjects. The variables manipulated are called independent variables or factors, while the variables measured are dependent variables.

Besides the conceptual definitions, variables can also be defined operationally. This is how the variable will be measured in practice. The measurement here is a procedure for assigning symbols, letters, or numbers to empirical properties of variables according to rules.

There are various levels of measurement and these levels differ according to how closely they approach the structure of the number system we use. It is important to understand these levels of measurement of variables in research, because knowing the levels of measurement helps you decide how to interpret the data from that variable. Further, knowing the levels of measurement helps you decide what statistical analysis is appropriate on the values that were assigned. Typically, there are four levels of measurement that are commonly distinguished in language studies. They are the nominal, ordinal, interval and ratio level of measurement.

1.4.1 Nominal level of measurement

A nominal level of measurement uses symbols to classify observations into categories that must be both mutually exhaustive and exclusive. "Mutually exhaustive" means that there must be enough categories that all the observations will fall into. "Mutually exclusive" means that the categories must be distinct enough that no observations will fall into more than one category. The nominal level of measurement is the most basic level of measurement: it is essentially labeling. In nominal level of measurement, all observations in one category are alike on some property, and they differ from the observations in the other category (or categories) on that property. For example, nouns can be sorted into countable nouns and uncountable nouns; the variable of perfect aspect may be measured by three categories: the past perfect, present perfect and future perfect. These categories must each be defined so that all possible observations will fit into one category but no more than one.

1.4.2 Ordinal level of measurement

An ordinal level of measurement uses symbols to classify observations into categories that are not only mutually exclusive and exhaustive. In addition, the categories have some explicit relationship among them. There is an ordering (or rank-ordered property) of categories, with one category better or worse, more or less than another. For example, observations of scores on a test can be ranked as categories of Excellent, Good, Fair, Failure, etc. However, no matter what kind of order it is, the distance between attributes does not have any meaning. For example, an experiment is performed to test the effect of certain linguistic features on the politeness of two sentences in a particular social context. Fifteen subjects are asked to