

Statistics in Linguistics

统计学在语言学中的应用 [英]

Christopher Butler

Basil Blackwell
World Publishing Corp

Statistics in Linguistics

Christopher Butler

Basil Blackwell
World Publishing Corp

© C. S. Butler 1985

First published 1985

Basil Blackwell Ltd
108 Cowley Road, Oxford OX4 1JF, UK

Basil Blackwell Inc.
432 Park Avenue South, Suite 1505,
New York, NY 10016, USA

All rights reserved. Except for the quotation of short passages for the purposes of criticism and review, no part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher.

Except in the United States of America, this book is sold subject to the condition that it shall not, by way of trade or otherwise, be lent, re-sold, hired out, or otherwise circulated without the publisher's prior consent in any form of binding or cover other than that in which it is published and without a similar condition including this condition being imposed on the subsequent purchaser.

British Library Cataloguing in Publication Data

Butler, Christopher
Statistics in linguistics.
1. Linguistics – Statistical methods
I. Title
410'.1'5195 P138.5
ISBN 0-631-14264-9
ISBN 0-631-14265-7 Pbk

Library of Congress Cataloging in Publication Data

Butler, Christopher
Statistics in linguistics.
Bibliography: p.
Includes index.
1. Linguistics – Statistical methods. I. Title.
P138.5.B87 1985 410'.2'8 85-3885
ISBN 0-631-14264-9
ISBN 0-631-14265-7 (pbk.)

Reprint authorized by Basil Blackwell
Reprinted by World Publishing Corporation, Beijing, 1991
Licensed for sale in China only
(excluding Hong Kong, Macao and Taiwan)
ISBN 7-5062-1055-X

Preface

In language study, as in the natural sciences, sociology or psychology, many kinds of work require the collection of quantitative data. The literary stylistician may wish to count the relative numbers of various colour terms, tense forms, alliterative sounds or some other linguistic feature of the texts in which he is interested. The language teacher or course designer may wish to obtain and compare measures of students' performances under two teaching methods. The theoretician may wish to count how many words in a corpus of texts occur once, how many twice and so on, and to compare these observed data with those predicted by a theoretical model of vocabulary distribution in texts. These are just a few examples of the many possible kinds of quantitative investigation into language. In all of them, we need ways of making sense of the data, and this is the purpose of statistical methods.

In many quantitative studies, we cannot investigate every possible example of the phenomenon we are interested in. In some cases exhaustive investigation is *theoretically* impossible; for example, if we were studying the time taken by informants to utter a particular sentence, the number of possible readings is infinite. In other cases, exhaustive examination is theoretically possible but impracticable; for instance, if we were examining some phonological feature of the English spoken in Birmingham, we could in theory obtain data from every Birmingham resident (or, better, from every resident satisfying a set of predetermined criteria for qualifying as a 'Birmingham speaker'); but this would be extremely time-consuming and difficult to organise, so that we should almost certainly be content with a *sample* from the total population we are concerned with. One important part of

statistics is concerned with methods of sampling, and with the relationships between measurements made on samples, and the properties of the populations these samples are intended to represent.

Once we have a set of data, either for every occurrence of our chosen phenomenon or for a sample of it, we usually need to summarise it in such a way that we can discern its general characteristics. The tools available for this task constitute *descriptive statistics*. Presented with a long list of numbers representing our observations, it is often not easy to see, at a glance, any general trends in the data. Such trends become more obvious when we look at the distribution of the data. For instance, in a language proficiency test on 100 learners, we might record marks out of 20. We can determine how many learners score 0, how many score 1, how many 2 and so on, up to 20, thereby drawing up a *frequency distribution* for the data, which may be converted to graphical form, and which gives an indication of the most typical score as well as the spread of marks. More precise measures of these properties can be obtained by performing certain statistical calculations on the data.

Very often, we are concerned not with the characteristics of just one set of data, but with the comparison of two (or more) sets. For example, we might be interested in testing the hypothesis that the performance of two groups of learners, taught by different methods, will differ in a language proficiency test; or we may wish to investigate whether the proportions of two pronunciations differ in the casual and formal speech of informants. In such cases we face the problem of designing our study in such a way that it will isolate just those phenomena we wish to test. The samples we use must be chosen so as to minimise variation arising from unwanted complicating factors, so that we can be reasonably confident that any effects owing to our chosen phenomenon are not swamped by other, 'irrelevant', effects. Experimental design is, or should be, inseparable from statistical work: no amount of sophisticated statistics can compensate for a badly designed investigation.

Where comparisons are involved, we need to know not only the general characteristics of each sample (such as the most typical value and the spread of values) but also whether the characteristics of the two samples are sufficiently different for us to conclude that there is a real effect that is due to the factor we are investigating. We can never be absolutely sure that a difference between

two sets of observations has not arisen 'by chance', owing to inherent variability in our material. We can, however, carry out tests which may allow us to claim 'real' differences with a specific margin of error, say 5 per cent, or 1 per cent, or 0.1 per cent. That is, we may, as a result of our calculations, claim to be 95 per cent sure, or 99 per cent sure, or even 99.9 per cent sure, that we have found a 'real' difference. This area, known as *hypothesis testing*, is an important part of *inferential statistics*.

In summary, then, whenever we wish to collect quantitative data on language, we need to pay careful attention to the design of our study, and to the selection of appropriate statistical methods for summarising the data, and for testing hypotheses concerning differences between sets of data. All these aspects of statistics are discussed in this book. However, since the book is introductory in scope, some techniques of interest to linguists, such as multiple correlation and regression, cluster analysis, and analysis of variance with more than one independent variable, are excluded. In order to deal adequately with these more advanced techniques, at least one further volume would be required.

Many courses on applications of statistics concentrate far too heavily on the methods themselves, and do not pay sufficient attention to the reasoning behind the choice of particular methods. I have tried to avoid this pitfall by discussing the 'why' as well as the 'how' of statistics. A difficult problem for the writer of any text on mathematical topics for non-mathematicians is how far to go into the derivation of formulae. While recognising that most linguists (including myself) will have neither an interest in the more theoretical side of the subject nor the mathematical background necessary for a full discussion, I feel that it is highly unsatisfactory for readers or students simply to be presented with a formula, with no explanation whatever of how it is arrived at. Where I thought it appropriate, I have attempted to give an idea of the rationale behind the various methods discussed in the book. Nevertheless, readers should find that their school arithmetic and algebra will see them through quite easily.

I should like to express my thanks to the various groups of students who have worked through the material presented here, and to Tim Gibson, who checked many of the exercises. My thanks go also to the following, for permission to use copyright or unpublished material:

Statistical tables in appendix 1: Dr H. R. Neave and his publishers George Allen & Unwin, for original or adapted versions of tables

2.1(a), 3.1, 3.2, 3.3, 5.1, 5.3, 6.2, 6.4 and 7.1 from *Statistics Tables for Mathematicians, Engineers, Economists and the Behavioural and Management Sciences* (1978); question 7 of chapter 9 exercises: Dr J. Connolly, for data from his article, 'Quantitative analysis of syntactic change', *Nottingham Linguistic Circular* 8/2, 108-18 (1979); question 6 of chapter 9 exercises: Dr J. Coates and Professor G. Leech, for data from their article, 'The meanings of the modals in modern British and American English', *York Papers in Linguistics* 8, 23-34 (1980); question 4 of chapter 11 exercises: Professor G. Wells and the publishers of *Research in Education*, for data from the article, 'Language use and educational success: a response to Joan Tough's *The Development of Meaning* (1977)', *Research in Education* 18, 9-34; questions 2 and 3 of chapter 2 exercises, question 3 of chapter 3 exercises, question 5 of chapter 9 exercises, questions 1 and 3 of chapter 10 exercises: Dr A. S. Crompton, for data from his work on stress and pause in French.

I should also like to thank Professor D. Crystal for advice on the production of the book.

Contents

Preface	vii
1 Some fundamental concepts in statistics	1
2 Frequency distributions	14
3 Measures of central tendency and variability	27
4 The normal distribution	44
5 Sample statistics and population parameters: estimation	53
6 Project design and hypothesis testing: basic principles	65
7 Parametric tests of significance	78
8 Some useful non-parametric tests	98
9 The chi-square test	112
10 The F distribution and its uses	127
11 Correlation	137
12 Statistics, computers and calculators	154
Appendix 1 Statistical tables	168
Appendix 2 Answers to exercises	182
References	211
Index	212

1 Some fundamental concepts in statistics

1.1 Introduction

In this chapter, some ideas of fundamental importance in statistics are introduced. We first consider the concepts of population and sample. We then differentiate between the two broad areas of statistical investigation: descriptive and inferential statistics. Finally, we consider some important ways in which statistical variables can be classified.

1.2 Populations and samples

In everyday speech, the term 'population' is normally taken to mean a collection of human, or at least animate, entities, as in 'the population of Greater London', or 'the rat population of the London sewers'. In statistics, however, the term is used more generally, to refer to any collection of entities, of whatever kind, that is the object of investigation. Thus, we may speak of the population of words in a given text, the population of nouns in Dickens's *David Copperfield* or the population of measurements of the time taken to utter a particular sentence on different occasions. It is with the characteristics of populations, or aggregates of individual entities, that statistics is most fundamentally concerned.

We may draw a distinction between *finite* and *infinite* populations. The population of nouns in *David Copperfield* or of past historic verbs in Camus's *La Peste* is finite: the number of entities is fixed and countable. Other kinds of population, however, are potentially infinite: in theory, at least, we could repeat our meas-

2 *Some fundamental concepts in statistics*

urement of utterance times for a sentence 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 our population is potentially infinite, or if it is finite but very large, we shall have to be content with *samples* drawn from the population concerned. The use of samples, even for the study of finite populations, cuts down the labour and cost involved in obtaining results, and minimises the errors which can easily be made during the processing of large amounts of statistical data. Let us imagine that we are interested in the acceptability judgements of native speakers of Parisian French when presented with a series of French sentences. Clearly, it would be impossible, in practice, to use every speaker of Parisian French as an informant (quite apart from the considerable difficulties involved in deciding what counts as 'Parisian French'). Instead, we take a sample of the speakers concerned, in the hope that our results will be generalisable to the population of such speakers as a whole.

Clearly, great care must be exercised in selecting samples if generalisation to the population is to be valid. Statistical methods of inferring the properties of populations from those of samples are based on the assumption that the samples are *random*. This does not mean that they are chosen in a haphazard manner: rather, it means that every unit in the population has an equal chance of being represented in the sample.

In order to select a truly random sample from a finite population, we need to have a complete list of the units in the population (often called the 'sampling frame'). Each unit in the sampling frame is allocated a number. One way of doing this is to put into a box a set of paper slips, each bearing a number, from 1 to the total size of the population, mix up the slips thoroughly, and withdraw a number of slips equal to the desired sample size. This method, though simple, is tedious, and a rather better way is to use a table of random numbers, in which the digits 0 to 9 are arranged in a very long sequence, in random order. Sets of statistical tables incorporating such information are readily available, and an excerpt from one such table is given as table A1 of appendix 1. The digits are grouped into fives, but this is merely for convenience in reading. Let us imagine that we wish to draw a sample of 50 lines of poetry from a text containing 800 lines. Since our population size is 800, a three-digit number, we need to take

groups of three digits from the table of random numbers. We may begin anywhere in the table; for convenience, let us start with the first three digits in column 1 of table A1. Reading down the column, we select items 494, 294, 252, 24, 694, 772 and 528 as the first seven 'members' of our sample. The next two figures are 987 and 850, which are too large for our population of 800, so we ignore them. The next members of the sample are lines 177, 637, 616, 422, 452, 403, 540, 491, 76, and so on.

A variation of simple random sampling is known as *systematic* or *quasi-random* sampling. Here, we decide the first unit in the sample by truly random methods, and then take units at equal intervals throughout the numbered population, the intervals being chosen to give the desired number in the final sample. Let us consider again the selection of a sample of 50 lines of poetry from a text of 800 lines. Our sampling interval will be $800/50$, or 16. We choose some point in the random number tables, and find the first two-digit combination which is smaller than or equal to 16. If this was, say, 12, we should take lines 12, 28, 44, 60, 76, and so on, up to and including line 796. Such a sample is not truly random, because the second and subsequent units are not chosen independently of the first. This does not matter seriously, provided that there is no periodicity in the population such that units with certain properties tend to recur at regular intervals. For instance, if the work of a particular prose writer showed a rather regular sentence length of about 20 words, and the sampling interval for selection of a sample of words from his work was also 20, the sample might contain a preponderance of words from the beginnings or ends of sentences, and this might seriously affect the representativeness of the sample.

A variation within systematic sampling is *block sampling*, where the position of the beginning of the sample is determined randomly, but the next N items are taken, where N is the size of the sample. For instance, in selecting 500 words from a text of 10 000 words, one might start at word 4581 (determined from random number tables), and take words 4581 to 5080 inclusive. The chances of such a passage being unrepresentative are considerable, and obviously vary according to the nature of the population and its structure.

It is important to realise that the selection of a sample by methods designed to achieve true randomness (such as using a table of random numbers) does not guarantee that the sample

4 *Some fundamental concepts in statistics*

arrived at will indeed be representative of the population. Consider, for example, the selection of a sample of ten people from the adult population of a village for linguistic testing, and assume, for simplicity, that there are equal numbers of men and women in the village. To calculate the chances of getting various possible mixes of men and women in the sample, we must make a short digression into elementary probability theory.

To simplify matters still further, let us first consider the selection of just three people from the village adult population. Since there are equal numbers of men and women, the probability that the first person selected will be a man is 0.5 (as is also, of course, the probability that this person will be a woman). Now, the laws of probability state that the probability of two independent events both occurring is equal to the product of the probabilities of the individual events. So the chances of both the first and the second members of our sample being men (or of both being women) is 0.5×0.5 , or 0.25, and the probability of all three being men (or all being women) is $0.5 \times 0.5 \times 0.5$, or 0.125. Now let us look at the possibility of getting two men and one woman in our sample. There are three ways in which this could happen:

Sample member no.	1	2	3
	man	woman	man
	man	man	woman
	woman	man	man

The probability of each of these outcomes is 0.125, so that the total probability of getting two men and one woman is 3×0.125 , or 0.375. Similarly, we can get two women and one man in any of three ways:

Sample member no.	1	2	3
	woman	man	woman
	woman	woman	man
	man	woman	woman

Again, the total probability is 0.375. We thus arrive at the situation shown in table 1.1. (Actually, this is an oversimplification, since the simple theory holds good only if each unit selected is put back into the population before the next choice is made, and this is clearly not possible in our situation, as we must avoid the possibility of the same person being chosen twice. This can, however, be ignored for our present purposes, provided that the

Table 1.1

Outcome	Probability
3 men	0.125
2 men + 1 woman	0.375
2 women + 1 man	0.375
3 women	0.125

population is large in relation to the sample size.) Note that the probabilities add up to 1.000, as indeed they must.

We could go through the same kind of reasoning for more complicated cases, but fortunately there is a general formula which will save us this trouble. If p and q are the probabilities of two mutually exclusive, independent events (such as choosing a man and choosing a woman, in our example), then the possible outcomes of n selections are determined by expanding the formula $(p + q)^n$. For the case where $n = 2$, we have

$$(p + q)^2 = p^2 + pq + pq + q^2 = p^2 + 2pq + q^2.$$

For $n = 3$, we have

$$\begin{aligned}(p + q)^3 &= (p + q)(p^2 + 2pq + q^2) \\ &= p^3 + 2p^2q + pq^2 + qp^2 + 2pq^2 + q^3 \\ &= p^3 + 3p^2q + 3pq^2 + q^3.\end{aligned}$$

For the case where $p = q = 0.5$, the terms evaluate to $0.125 + 0.375 + 0.375 + 0.125$, and this is the result we obtained from first principles above. This type of distribution is known as the *binomial distribution*. We shall not discuss it further here: details can be found in any comprehensive statistics textbook (see, for example, Ferguson 1981).

We may now return to our original problem: predicting the probabilities of various outcomes in the selection of ten people from our village population. If the probabilities are worked out for $n = 10$, using the expansion discussed above, we obtain the results shown in table 1.2.

The reader will remember that the point of this digression was to enable us to see just how representative of the population a

Table 1.2

Men	Women	Probability
0	10	0.001
1	9	0.010
2	8	0.044
3	7	0.117
4	6	0.205
5	5	0.246
6	4	0.205
7	3	0.117
8	2	0.044
9	1	0.010
10	0	0.001

randomly selected sample of ten people might be. Ideally, we should like the proportion of men and women in the sample to reflect that in the population: that is, five men and five women. We see from the above that the probability of just this outcome is about a quarter (to be more precise, 0.246). The chances of getting eight or more men in the sample (or, of course, eight or more women) are $0.044 + 0.010 + 0.001$, or 0.055 – that is, over 1 in 20. Putting this another way, if we took 20 samples of ten people from the village, we could expect one of these to contain eight or more men, and one to contain eight or more women. Such a sample would hardly be representative of the sex distribution of the village. Furthermore, it is not hard to see that any types of unit which have a small overall proportion in the population may well not be represented in the sample at all.

² One way to minimise this problem is to select a sample by ~~random~~ *stratified random sampling*. If the proportions of various subgroups in the population are known, then random sampling can be undertaken within each stratum, and the resulting subsamples combined to give an overall sample. For example, in selecting a sample of ten adults from our village, we might choose, by random methods, five men and five women, and combine the two subsamples. Such a procedure would be an example of *proportional stratified random sampling*, since the proportion of men and women in the final sample is the same as in the population. As a further example of proportional stratification, consider the problem of selecting a sample that will be as representative as possible

Table 1.3

	Early period	Middle period	Late period
Novels	500 000	700 000	300 000
Short stories	100 000	200 000	200 000

of the range of an author's literary prose works. Let us assume that the author's total output of 2 million words is divided into genres and periods as in table 1.3. We might then choose our sample so that $500\,000/2\,000\,000$, or 0.25 of it, came from early novels, $700\,000/2\,000\,000$, or 0.35, from middle novels, and likewise 0.15 from late novels; 0.05 from early short stories, 0.1 from middle-period short stories, and 0.1 from late short stories. Within each of these categories, we could select our subsample by simple random sampling, or by systematic or even block sampling methods.

A stratified framework has an additional advantage in that it allows comparisons to be made between subsamples corresponding to the various groupings. This is possible with a proportionally stratified sample such as those discussed above. Often, however, the optimum situation for comparison of subgroups is one in which these subgroups are of equal size, despite their unequal proportions in the population as a whole. A sample chosen in this way is said to be a *disproportionally* stratified sample. Imagine, for instance, that we wish to select a sample of 120 000 words from the 2 million word output of our hypothetical author. Table 1.4 compares the number of words taken from each subgroup under (1) proportional and (2) disproportional stratified sampling.

If we wish to use the results from a disproportionally stratified sample to estimate the properties of the population as a whole, we must obviously give the various strata different weightings. For instance, there are only two-thirds as many words from early novels in the disproportional sample as in the proportional sample, so that we should weight by a factor of $3/2$, or 1.5, the contribution of the results from this subgroup; and so on.

We should also mention *multi-stage sampling*, which, as the term suggests, consists of sampling by successive stages, treating the sample at one stage as the population for the next. For example, if we wished to obtain a sample of 100 pages from the

8 *Some fundamental concepts in statistics*

Table 1.4

	<i>Proportional</i>	<i>Disproportional</i>
Novels		
Early	30 000	20 000
Middle	42 000	20 000
Late	18 000	20 000
Short stories		
Early	6 000	20 000
Middle	12 000	20 000
Late	12 000	20 000

novels of Dickens, we might first select five novels by a random procedure, using the total set of novels as the population, and then within each novel select 20 pages, again at random. Sometimes, investigations which at first sight look as if they are based on single-stage sampling are in fact better regarded as multi-stage. Consider the situation where we wish to select two samples from a class of language learners, in order to compare the effectiveness of two different teaching methods. The samples chosen for comparison are samples of the class taken as a population; but we should probably want to try to generalise from that particular class of learners to the whole population of learners of similar characteristics, at a similar stage of development.

This last point brings us to an important practical issue. We have seen that true random sampling from a finite population requires the listing of the entire population. It is often impossible, or at least impracticable, to obtain a list of the whole population in which we are interested. In some cases of sampling involving human populations, electoral and similar lists can be useful; but even here, we must be careful not to use, as a sampling frame, a list that has some in-built bias which might prejudice our results. A list of telephone subscribers, for example, would almost certainly bias our sample towards the middle and upper ranges of the social class scale. In many cases, we have to make do with a 'population' which is 'given' by some practical circumstance, such as the members of a school, or even a class within a school. In such cases, if we wish to generalise beyond that group we must, if possible, do our best to demonstrate that the properties of the group do not differ radically from those of the population in which we are really interested. For instance, we might try to show

that the distribution of ages, sexes, IQ ratings and social classes in a particular school is not too far removed from the norm for schoolchildren in the country as a whole. Often, however, the norm itself is unknown and difficult to determine; furthermore, it may not always be clear just what factors are relevant. We must therefore be on our guard against the understandable temptation to claim greater generalisability for our results than is warranted by the situation.

Finally in this section, we must introduce two more technical terms connected with populations and samples. The properties of populations are normally called *parameters*, while the properties of samples from populations are called *statistics* (or sometimes *estimates*). Imagine that we select 1 000 words randomly from a text of 50 000 words, and find that their mean (a kind of 'average') length is 3.54 letters. This measure is a statistic for the sample, and is used as an estimate of the true population parameter, which probably will not be exactly 3.54 letters. Fortunately, as we shall see, methods are available for assessing the degree of confidence we may have in the reliability of such estimates. Statistics are normally symbolised by Roman letters (for instance, $\bar{x}$, read as 'x-bar', for the mean of a sample), while parameters have symbols consisting of Greek letters (such as μ , 'mu', for the mean of a population).

1.3 The descriptive and inferential functions of statistics

In the preface to this book, a distinction was made between the *descriptive* and *inferential* functions of statistics. We shall now examine this distinction in rather more detail, by means of a practical illustration.

Let us suppose that we have selected two samples of 30 language learners each, one taught by traditional methods, the other by means of the language laboratory. We have given both groups a proficiency test in the target language, with the following results:

Class A (language laboratory): marks out of 20:

15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17, 13, 14, 13, 15, 17, 19, 17, 18, 16, 14.

Class B (traditional): marks out of 20:

11, 16, 14, 18, 6, 8, 9, 14, 12, 12, 10, 15, 12, 9, 13, 16, 17, 12, 8, 7, 15, 5, 14, 13, 13, 12, 11, 13, 11, 7.