THE FUNDAMENTALS OF STATISTICS

BY

L. L. THURSTONE, M.E., PH.D.
Associate Professor of Psychology,
University of Chicago

New York
THE MACMILLAN COMPANY
1925

PRINTED IN THE UNITED STATES OF AMERICA

Set up and electrotyped. Published February, 1925. Reprinted December, 1925.

Norwood Press
J. S. Cushing Co. — Berwick & Smith Co.
Norwood, Mass., U.S.A.

Editor's Introduction

IT is impossible for any person to read very much of present-day educational literature with pleasure and understanding unless he is acquainted to some extent with the method and the terminology employed in conducting and presenting the results of statistical investigation. The character of educational literature has undergone a fundamental change during the past ten or fifteen years. Doubtless some of the readers of these lines can remember the time when articles and books relating to educational values, methods, or administration rarely, if ever, contained tables of data or graphs showing the distribution of the facts bearing upon any phase of educational procedure. In those days a teacher hardly ever — probably never — came across in her professional reading such terms as *correlation coefficient*, *median*, *mode*, *variability*, *frequency tables*, *frequency surface*, *standard deviation*, *probability curve*, *percentile ranks*, and so on. But choose at random twenty-five books or articles on educational subjects that have recently appeared and that are regarded as up-to-date, and it will undoubtedly be found that the content of twenty of them is based upon investigations involving statistical method, and that the conclusions are phrased in terms which can be compre-

hended only when one is familiar with statistical modes of organizing and presenting data.

The time is passing rapidly when one can safely discuss most educational subjects without having first made a statistical investigation of the topics which he treats. Formerly educational writers discussed problems of values, methods, and administration without any attempt to be statistically sound and consistent in their premises or conclusions. They relied upon "reason" to guide them accurately in matters pertaining to educational procedure. They analyzed problems under consideration, and on the basis of experience, logic, or "common sense" they deduced principles without deeming it necessary to determine to what extent these principles would hold good in the concrete situations to which they related. But to-day we have slight confidence in the validity or value of principles arrived at in this fashion. We demand of anyone who presumes to speak authoritatively upon educational themes that he shall first have gathered the data pertaining to the themes in an accurate way, and that he shall then have treated the data so as to clarify obscure matters, eliminate error from his conclusions, and show how his facts are distributed.

It is universally recognized that all the data relating to educational problems are complex; the situations out of which they have sprung are complicated, and in general any factor pertaining to any aspect of education is intimately bound up functionally with other factors, so that it is not possible to de-

scribe the characteristics or measure the force of the factor in question unless it can be isolated for purposes of measurement from the factors with which it is ordinarily associated. The untrained student cannot accomplish this in treating the problems which he wishes to investigate. Even if he can secure data by proper modes of procedure, he does not know how to organize the data and present conclusions so as to reveal their trend and meaning. For this reason it is essential that, before attempting investigation, he should become familiar with methods of procedure that will enable him to avoid error in the treatment of his data and to derive principles which the data should yield.

Unfortunately the books that deal with statistical method in the treatment of educational data have seemed to the novitiate in educational investigation to be technical and forbidding. The very appearance of most books on statistics has been enough to deter the novice from attempting to master this subject. The present writer believes that Mr. Thurstone's volume will be more acceptable than the typical textbook on statistical method to those who wish to become familiar with the method so that they can read educational literature understandingly and agreeably, and so that they may participate according to their resources in the investigation of educational problems. The author has succeeded in discussing every detail of statistical procedure in such a simple and clear way and in such terminology that it can be comprehended by a teacher or investigator who has not

been able to devote much time and energy to statistical study. Every essential principle is illustrated by concrete, impressive instances. The book might be regarded as a primer or introduction to the fundamental principles of statistics. The author has had extensive experience in teaching statistical method to students in educational psychology. He has made himself familiar with the difficulties which novices encounter in understanding the logical implications of statistics, and he has acquired exceptional skill in making these implications clear and intelligible.

The book may be heartily commended for its simplicity and clarity to all who wish to contribute to educational investigation as well as to those who wish to understand the results of investigations made by others.

M. V. O'Shea.

The University of Wisconsin.

Preface

THIS textbook in statistics is the result of seven years of teaching the fundamental principles of statistics and mental measurements to classes of about thirty graduate students annually. I have found that the majority of the students who enter psychology in their graduate work come into it from undergraduate majors in economics, literature, languages, psychology, and other unmathematical subjects. They have dodged the college course in mathematics, and they seldom have any occasion to keep fresh their high school mathematics. For this reason it has been necessary to assume on their part very little knowledge of even the fundamentals of high school algebra, but this lack is often compensated for by a keen critical ability in the logic of the subject. This critical attitude toward statistical work I have always tried to encourage, and I have discouraged, as far as possible, the blind substitution of numbers into formulæ. It happens often that a student learns readily to calculate correlation coefficients and to draw pretty charts, but unless he understands the logical implications of his statistical work, more harm than good has been done. It has been my aim throughout these lessons to make clear the meaning and implications of statistical procedures.

I owe to my students much of the form of the explanations, since I have included in these lessons those particular examples, methods, and teaching tricks which have seemed to be most successful in explaining the subject to unmathematically inclined students.

I have borrowed extensively from various textbooks on statistics. The correlation data sheet is a modification of Thorndike's correlation table. I wish to acknowledge the use of *Figures 30* and *38* and some of the related text on percentiles and on the correlation table from my articles in the *Journal of Educational Research*, for which the publishers have kindly given their consent. I have adapted *Tables 19* and *20* from the extensive tables prepared by W. F. Sheppard and published in "Tables for Statisticians and Biometricians," edited by Karl Pearson. The two tables are brief, but they should serve the purposes of the beginner in statistical work. He will refer to Sheppard's tables for more accurate determinations.

I hope that this manual may prove useful not only to the beginning student in statistics but also to those workers in the field of mental measurement who do not feel at home in the logical interpretation of their statistical work. Many of the outlines and sample problems have been so arranged that they should prove helpful also for reference purposes. After the student has become familiar with the logical aspects of the measures of central tendency and variability and the meaning of the correlation table, he should be able to pursue his statistical studies further in the

more comprehensive textbooks, such as those of Yule, Bowley, Elderton, Brown and Thompson, and Kelley. The one outstanding textbook in statistics for the student whose training in mathematics is limited is that of Yule. It is so exceedingly well done that it would be folly for me to attempt anything more than an introduction to it.

Teachers and students who may use this manual will confer a favor, for which I shall be grateful, if they will call my attention to errors either in the text or in the arithmetical work.

L. L. THURSTONE.

CHICAGO, July, 1924.

Contents

List of Tables

List of Figures

The Fundamentals of Statistics

Chapter One

The Frequency Table

When we have a collection of facts in numerical form, the first statistical task is usually to classify the data in some way. Suppose that a mental test has been given to three hundred students and that the papers have been scored. Some one inquires for the score of Mr. Jones, and we find that his score is 79. Is that a high score or a low score? We cannot answer that unless we know how many of the students scored *above* 79 and how many of them scored *below* 79. If all the other students scored below 79, then Jones' score is high; but if all the other students scored above 79, then Jones' score is low. If 150 of the 300 students scored above 79 and 150 of them below 79, then Jones has an average or ordinary score. This will suffice to show that it is not enough to give the test and score the papers; we must also classify the scores so that we may know how many students scored in the nineties, how many in the eighties, and so on. Such a table is called a *frequency table.*

An intelligence test was given to the freshmen at Swarthmore College. In *Table 1* we have a list of scores. Each number is the score of a student. If we want to know how many students obtained scores above 79, it is necessary to look through the whole

62	129	95	123	81	93	105	95	96	80
123	60	72	86	108	120	57	113	65	108
109	84	121	60	84	128	100	72	119	103
77	91	51	100	63	107	76	82	110	63
104	107	63	117	116	86	115	62	122	92
69	116	82	95	72	121	52	80	100	85
94	84	123	42	90	91	81	116	73	79
100	79	101	98	110	95	67	77	91	95
79	92	73	83	74	125	101	82	71	75
125	56	86	98	106	72	117	89	99	86
87	90	80	131	102	117	98	74	101	82
110	137	99	65	113	85	82	90	102	57
139	74	149	114	74	102	69	134	78	106
75	106	85	103	78	106	102	94	108	90

Table 1. Scores of Swarthmore College freshmen in an intelligence test

table. That is not necessary when the data are arranged in the form of a frequency table, as shown in *Figure 1*.

The frequency table is prepared as follows:

1. Arrange a *data sheet*[1] with the three headings *Score*, *Tabulation*, and *Frequency*, as shown in *Figure 1*.

2. Read off the scores in *Table 1* and for each one record a check mark as shown in *Figure 1*. The subsequent counting is facilitated if every fifth mark is made slanting across the preceding four checks, as shown in *Figure 1*.

[1] A data sheet is a sheet ruled with vertical columns for recording numerical or other data. In recording facts on a data sheet, be sure to label each column.

3. Add the check marks in each *row* and record the sums under *Frequency*.

Score	Tabulation	Frequency
0 – 9		
10 - 19		
20 - 29		
30 - 39		
40 - 49	𝍸	1
50 - 59	𝍸𝍸	5
60 - 69	𝍸𝍸 𝍸𝍸 ||	12
70 - 79	𝍸𝍸 𝍸𝍸 𝍸𝍸 𝍸𝍸 |	21
80 - 89	𝍸𝍸 𝍸𝍸 𝍸𝍸 𝍸𝍸 |||	23
90 - 99	𝍸𝍸 𝍸𝍸 𝍸𝍸 𝍸𝍸 |||	23
100 - 109	𝍸𝍸 𝍸𝍸 𝍸𝍸 𝍸𝍸 𝍸𝍸	25
110 - 119	𝍸𝍸 𝍸𝍸 ||||	14
120 - 129	𝍸𝍸 𝍸𝍸 |	11
130 - 139	||||	4
140 - 149	|	1
150 - 159		
160 - 169		
Total number of students =		140

Figure 1. Frequency table of scores in intelligence test for Swarthmore freshmen

4. Add the frequency *column*. This sum is the total number of cases and should agree with the number of scores in *Table 1*.

It is now possible to answer such questions as these:

1. How many students obtained scores between 70 and 79? (21)
2. How many students obtained scores between 60 and 69? (12)
3. How many students obtained scores above 99? (55)
4. How many students obtained scores below 70? (18)
5. How many students obtained scores between 20 and 49? (1)
6. What per cent of the freshman class obtained scores between 80 and 89? (16%)
7. What per cent of the freshman class obtained scores between 40 and 79? (28%)
8. Is a score of 95 high, average, or low? (Average)
9. What per cent of the class exceeded the score of 89? (56%)

All these questions and others of the same kind can be answered by referring to the frequency table.

A *variable* is any quantity which can have different numerical values. It is any varying quantity. Examples of variables are ages, birth- and death-rates, prices, wages, barometer readings, rainfall records, and city populations. The scores obtained in intelligence tests constitute a variable. We may consider as a variable the scores made by the different persons in a group. We may also consider as a variable the scores that a single person makes on the same test on different occasions.

Variables may be classified as *continuous* and *discontinuous*. If it is possible for a variable to change its numerical value by infinitesimally small degrees, it is called a continuous variable. If this is not possible, the variable is called discontinuous. Temperature, for instance, changes from 68 degrees to a new value, such as 69 degrees, by passing through all the intermediate values. Therefore temperature is considered a continuous variable. The number of freight cars in a train is a discontinuous variable because this variable does not change by passing through all intermediate values. If the train is 68 cars long, it cannot be made 69 cars long by passing through the intermediate values of, let us say, $68\frac{1}{3}$ cars and $68\frac{5}{7}$ cars.

The *range* is the difference between the maximum and minimum values of the variable in any series. In this group of Swarthmore freshmen the range of intelligence test scores is 107 because this is the difference between the highest and lowest scores in the freshman class (149 and 42 respectively).

The *class interval* is one of the equal parts into which a scale is divided for convenience in tabulation. If we were tabulating the ages of employees, we should probably classify them by years. In this case the year would be the class interval. If we desired a more refined classification, we might classify their ages by half-years or by months. This is sometimes done in classifying the ages of children. If we were classifying people by their yearly salaries, we might select $100 as a convenient class interval. In