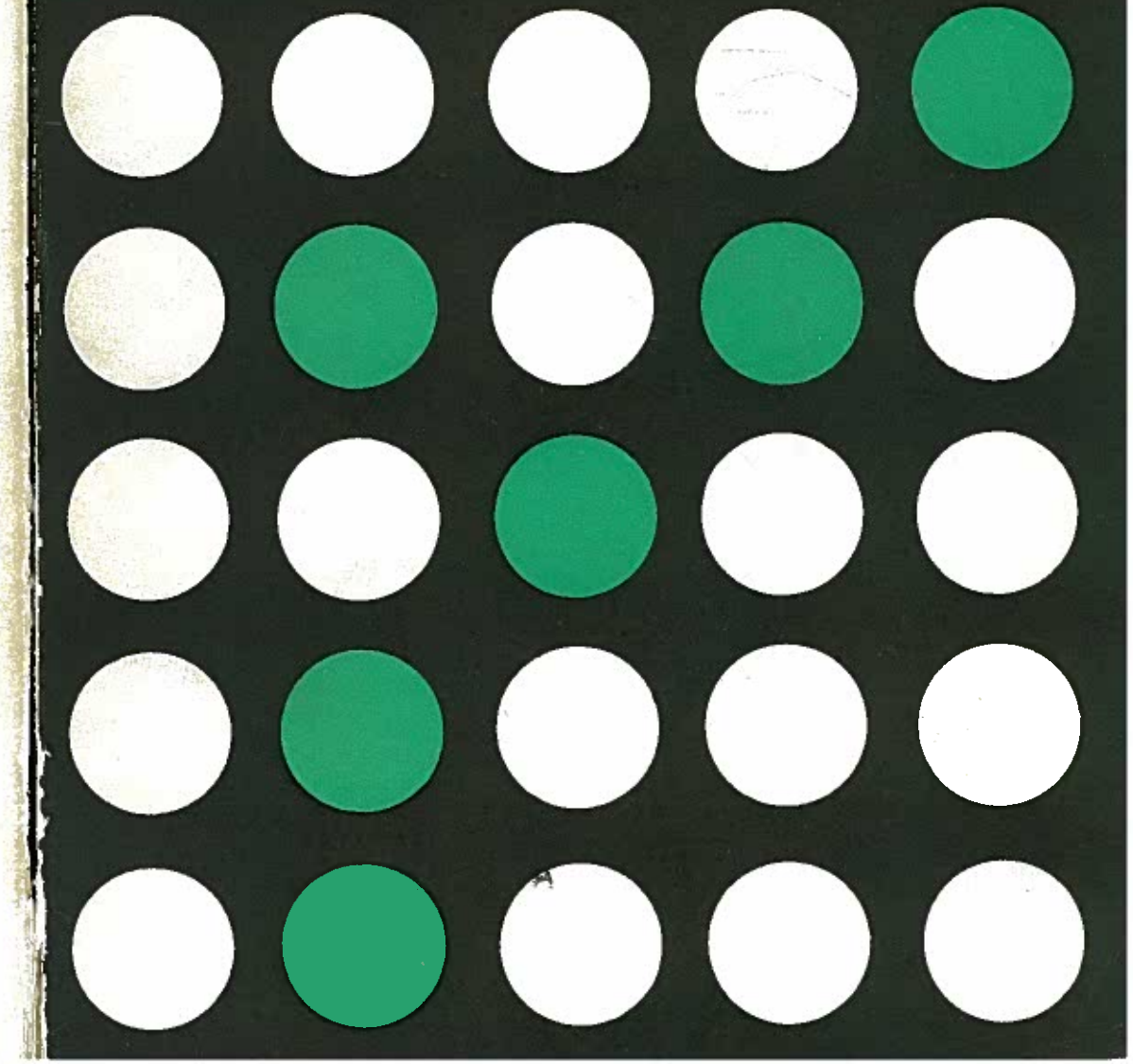
New Zealand	Government Printing Office: Government Bookshops at State Advances Buildings, Rutland Street, P.O. Box 5344, Auckland; 20 Molesworth Street, Private Bag, Wellington; 112 Gloucester Street, P.O. Box 1721, Christchurch; Stock Exchange Building, corner Water and Bond Streets, P.O. Box 1104, Dunedin.
Nigeria	University Bookshop Nigeria Ltd., University College, Ibadan.
Norway	Johan Grundt Tanum Forlag, Karl Johansgt. 43, Oslo.
Pakistan, West	Mirza Book Agency, 65 The Mall, Lahore 3.
Panama	Agencia Internacional de Publicaciones J. Menéndez, Apartado 2052, Panama.
Paraguay	Agencia de Librerías de Salvador Nizza, Calle Pte. Franco No. 39-43, Asunción.
Peru	Librería Internacional del Perú, S.A., Casilla 1417, Lima.
Philippines	The Modern Book Company, 518-520 Rizal Avenue, Manila.
Poland	Ars Polona, Krakowskie Przedmiescie 7, Warsaw.
Portugal	Livraria Bertrand, S.A.R.L., Rua Garrett 73-75, Lisbon.
South Africa	Van Schaik's Book Store Ltd., P.O. Box 724, Pretoria.
Spain	José Bosch Librero, Ronda Universidad 11, Barcelona; Librería Mundi-Prensa, Castelló 37, Madrid; Librería General, S. Miguel 4, Saragossa.
Sweden	C.E. Fritze, Fredsgatan 2, Stockholm 16; Gumperts A.B., Göteborg; Universitetsbokhandel, Sveavägen 166, Stockholm Va.
Switzerland	Librairie Payot, S.A., Lausanne and Geneva; Hans Raunhardt, Kirchgasse 17, Zurich 1.
Syria	Librairie Internationale, B.P. 2456, Damascus.
Taiwan	The World Book Company Ltd., 99 Chungking South Road, Section 1, Taipeh.
Thailand	Requests for FAO publications should be addressed to: FAO Regional Office for Asia and the Far East, Maliwan Mansion, Bangkok.
Turkey	Librairie Hachette, 469 Istaklal Caddesi, Beyoglu, Istanbul.
United Kingdom and Crown Colonies	H.M. Stationery Office, 49 High Holborn, London W.C.1; P.O. Box 569, London S.E.I. *Branches at*: 13a Castle Street, Edinburgh 2; 35 Smallbrook, Ringway, Birmingham 5; 50 Fairfax Street, Bristol 1; 39 King Street, Manchester 2; 109 St. Mary Street, Cardiff; 80 Chichester Street, Belfast.
United States of America	Columbia University Press, International Documents Service, 2960 Broadway, New York 27, New York.
Uruguay	Héctor d'Elía - Editorial Losada Uruguaya S.A., Colonia 1060, Montevideo.
Venezuela	Suma, S.A., Calle Real de Sabana Grande, Caracas; Librería Politécnica, Apartado del Este 4845, Caracas.
Yugoslavia	Drzavno Preduzece, Jugoslovenska Knjiga, Terazije 27/11, Belgrade; Cankarjeva Zalozba, P.O. Box 201 - IV, Ljubljana.
Other countries	Requests from countries where sales agents have not yet been appointed may be sent to: Distribution and Sales Section, Food and Agriculture Organization of the United Nations, Via delle Terme di Caracalla, Rome, Italy.

FAO publications are priced in U.S. dollars and pounds sterling. Payment to FAO sales agents may be made in local currencies.

Price: $6.00 or 30s. PM33003/9.66/E/1/3800

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS



QUALITY OF STATISTICAL DATA

by

S. S. ZARKOVICH

Chief, Methodology Branch

Statistics Division, FAO

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS

Rome 1966

Printed in Italy

PREFACE

The body of this book consists of the material presented by the author at various seminars and training centers organized by FAO. The aim of these lectures was to spread awareness of the quality problem of statistical data and to promote interest in quality checks as a source of guidance on the adequate uses of data and on the ways and means of improving the methods used.

The composition of the book reflects these aims. First of all, an effort was made to explain in some detail what happens if errors are introduced into the data collected and what problems arise as a result. Afterward examples and experiences were presented in order to illustrate the significance of these problems. Similarly, in the presentation of techniques that might be involved in checking the quality of data, emphasis was laid on the explanation of the logic of the procedures rather than on the analysis of various techniques that were actually used under specific circumstances. At the end of the book a bibliographical list has been added. This list may offer some useful guidance to those who want to continue the study of material covered in the book.

In writing the book an effort was made to give a text that would be as simple as possible. In fact, an elementary course of statistical theory and a basic course on the theory of sample surveys are sufficient for an understanding of the text.

The reader may be surprised to see in this book definitions of certain basic concepts of survey techniques, such as biased and unbiased estimates, mean square error, etc. In spite of the fact that these concepts are known to all statisticians, it was thought necessary to include them in a text that aims at dealing systematically with all the topics that are encountered in a study of the quality of data.

A draft of this book has appeared in mimeographed form under the title *Sampling methods and censuses*, Vol. II: *Quality of statistical data*, FAO, 1963. An earlier and very summarized version of it was presented at the annual meeting of the Yugoslav Statistical Society under the title *Quality problems of statistical data* (in Serbo-Croatian), Ljubljana, 1956.

P.V. SUKHATME
Director, Statistics Division
FAO, Rome

CONTENTS

1. SOME BASIC CONCEPTS

1.1 Definition of errors

Before a survey can be made there are many factors that have to be determined. Such factors are concepts and definitions, methods of collecting data, the units to be used in expressing the response, the tabulation program, the survey program, the wording of questions, etc. We refer to all these factors under the general term of the *adopted system of work*. Accordingly, the adopted system of work shows what data are to be collected in a survey, in what way they are to be collected, etc.

The adopted system of work is shaped according to the aims of the survey. Since it represents a fixed system of concepts, definitions, procedures, and operations that constitute the survey, the specification of the adopted system of work makes it possible to judge whether the action taken is in agreement with the action prescribed. Needless to say, this possibility is sometimes only theoretical.

On the basis of the concept of the adopted system of work the concept of the *true value* can be defined. The true value is simply the result that should be obtained in a particular survey operation if the adopted system of work is carried out correctly. The true value is the *ideal result* of a particular survey operation; it is obtained if the work is done in *absolute conformity* with the adopted system of work.

There are several types of true value. The first is the *individual true value* of a characteristic for a given unit of population. The individual true value follows from the application of the adopted system of work in obtaining the value of a characteristic for a given unit. If in a census of population the age of the head of the household is required in years completed at the last birthday, the true value of this particular item is the number of years the head has in fact completed irrespective of whether he is aware of this value or not, and independently of what he has stated in the census. The true value of the total area of a holding as expressed in hectares would be the sum of the true values of the area of individual fields rounded off to the nearest integer. It is therefore evident that, after

the adopted system of work is fixed, the true value becomes a defined quantity.

In some cases it will not be easy to visualize the meaning of true values. By way of example, it is sufficient to remind the reader of "intention" surveys, such as planting intentions. However, the difficulty arising here as well as the practical difficulty of ascertaining true values should not prevent the use of this concept, since, in fact, the real essence of errors in statistics can hardly be described without it.

In addition to individual true values we also speak of true values of totals, averages, proportions, coefficients of correlation, and other statistical measures. The meaning of these concepts is obvious.

In order to define the true value of the population total we use the symbol x_i to designate the true value of a characteristic for the i-th unit of the population. It is assumed that the total number of units of this population is equal to N. The true value of the population total for this characteristic is then defined as

$$X = \sum_{i}^{N} x_i \tag{1.1}$$

Definitions of the true values of other statistical measures are obvious.

It is clear that individual true values are not always achieved in survey practice for all the units. The results achieved factually will be called *survey values*. The survey value of the i-th unit of the population for the same characteristics as before, viz. x_i, will be designated by z_i. By analogy with the definitions of true values we distinguish *individual survey values* and survey values of various statistical measures. It is clear that

Individual errors may be positive and negative. If the survey value is equal to the corresponding true value, i.e., when $z_i = x_i$ or $d_i = 0$, we say that z_i is *accurate*. On the other hand, if $d_i \neq 0$, z_i will be called *inaccurate*.

The following are some additional concepts. From (1.3) we have

$$z_i = x_i + d_i \tag{1.4}$$

and

$$\sum_{i}^{N} z_i = \sum_{i}^{N} x_i + \sum_{i}^{N} d_i$$

or

$$Z = X + D \tag{1.5}$$

The quantity D is called the *bias*. Clearly, if $D = 0$, the survey value of the population total for a given characteristic is equal to the corresponding true value. In this case Z is said to be accurate or *unbiased*. Vice versa, if $D \neq 0$, Z is said to be *biased*.

Using (1.3) or (1.4) it is easy to define biases in other statistical measures.

From a practical point of view great importance is attached to the *frequency distribution of individual errors*. If positive and negative errors are distributed at random around zero, the estimates of totals and averages will be unbiased. In many cases, however, there is some pattern in errors in the sense that either positive or negative errors predominate. In such a case we speak of *systematic errors*. Totals and averages based on data subject to systematic errors will normally be biased. The bias is thus the *net effect* of all the errors.