

New Zealand

Nigeria  
Norway  
Pakistan, West  
Panama

Paraguay

Peru  
Philippines  
Poland  
Portugal  
South Africa  
Spain

Sweden

Switzerland

Syria  
Taiwan

Thailand

Turkey  
United Kingdom  
and Crown  
Colonies

United States  
of America  
Uruguay

Venezuela

Yugoslavia

Other countries

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.

University Bookshop Nigeria Ltd., University College, Ibadan.  
Johan Grundt Tanum Forlag, Karl Johansgt. 43, Oslo.  
Mirza Book Agency, 65 The Mall, Lahore 3.  
Agencia Internacional de Publicaciones J. Menéndez, Apartado 2052, Panama.

Agencia de Librerías de Salvador Nizza, Calle Pte. Franco No. 39-43, Asunción.

Librería Internacional del Perú, S.A., Casilla 1417, Lima.  
The Modern Book Company, 518-520 Rizal Avenue, Manila.  
Ars Polona, Krakowskie Przedmiescie 7, Warsaw.

Livraria Bertrand, S.A.R.L., Rua Garrett 73-75, Lisbon.  
Van Schaik's Book Store Ltd., P.O. Box 724, Pretoria.  
José Bosch Librero, Ronda Universidad 11, Barcelona; Librería Mundi-Prensa, Castelló 37, Madrid; Librería General, S. Miguel 4, Saragossa.

C.E. Fritze, Fredsgatan 2, Stockholm 16; Gumperts A.B., Göteborg; Universitetsbokhandel, Sveavägen 166, Stockholm Va.

Librairie Payot, S.A., Lausanne and Geneva; Hans Raunhardt, Kirchgasse 17, Zurich 1.

Librairie Internationale, B.P. 2456, Damascus.

The World Book Company Ltd., 99 Chungking South Road, Section 1, Taipeh.

Requests for FAO publications should be addressed to: FAO Regional Office for Asia and the Far East, Maliwan Mansion, Bangkok.

Librairie Hachette, 469 Istaklal Caddesi, Beyoglu, Istanbul.  
H.M. Stationery Office, 49 High Holborn, London W.C.1; P.O. Box 569, London S.E.1. 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.  
Columbia University Press, International Documents Service, 2960 Broadway, New York 27, New York.

Héctor d'Elia - Editorial Losada Uruguay S.A., Colonia 1060, Montevideo.

Suma, S.A., Calle Real de Sabana Grande, Caracas; Librería Politécnica, Apartado del Este 4845, Caracas.

Drzavno Preduzece, Jugoslovenska Knjiga, Terazije 27/11, Belgrade; Cankarjeva Založba, P.O. Box 201 - IV, Ljubljana.

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.

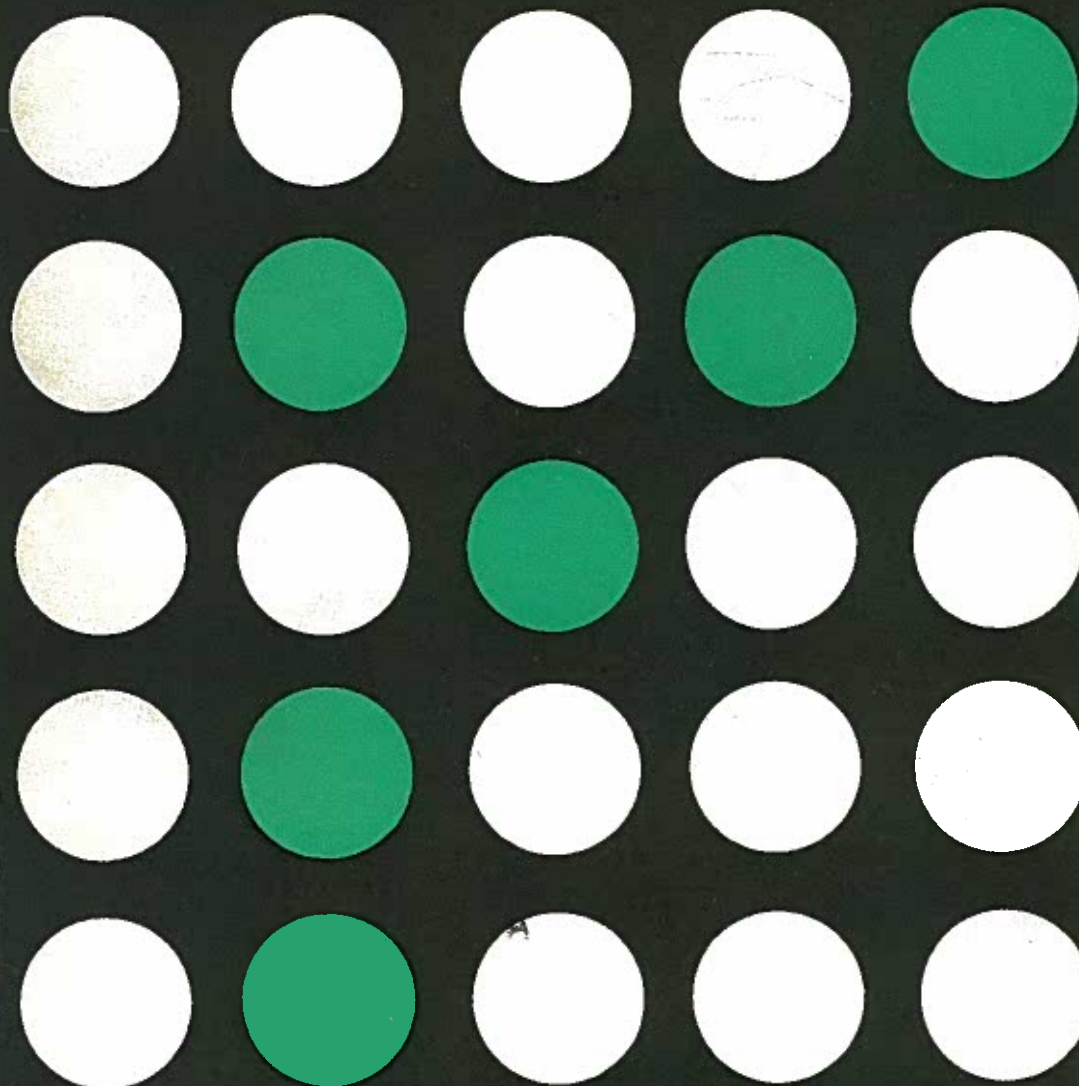
*FAO publications are priced in U.S. dollars and pounds sterling. Payment to FAO sales agents may be made in local currencies.*

Price: \$6.00 or 30s.

PM33003/9.66/E/1/3800

v 2  
1966  
cop. 2

## 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

311  
Z 111  
Eng. ed.  
v. 2  
1966  
Gp. 2

## 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

PREFACE	v
1. SOME BASIC CONCEPTS .....	1
1.1 Definition of errors .....	1
1.2 Where do errors appear? .....	4
1.3 Classification of errors .....	6
1.4 Relative character of errors .....	9
1.5 Biased procedure of estimation .....	10
1.6 Biased versus unbiased estimates.....	14
1.7 Illustrations .....	17
1.8 Further aspects of the presence of errors in data ...	20
1.9 Checking the quality of data .....	22
2. USE OF POST HOC TECHNIQUES IN CHECKING THE QUALITY OF DATA .....	24
2.1 Comparisons with data from independent sources ..	24
2.2 Consistency studies .....	28
2.3 Internal consistency .....	33
2.4 Cohort survival .....	34
2.5 Drawbacks of post hoc techniques .....	37
2.6 Advantages of sampling methods .....	38

3. MEANING OF QUALITY CHECKING .....	40
3.1 Assumption of random response variation .....	40
3.2 Unique nature of response .....	43
3.3 Difficulties of generalization .....	47
3.4 Accuracy checks .....	48
3.5 Approximation to true values .....	51
3.6 Supporting evidence .....	53
3.7 Other types of quality checks .....	55
3.8 Terminology .....	58
4. BIASED PROCEDURES .....	59
4.1 Definition of biased procedures .....	59
4.2 Measurement procedures .....	60
4.3 Selection procedure .....	64
4.4 Control of selection biases .....	73
4.5 Biased estimation procedure .....	76
5. BIASED TOOLS .....	82
5.1 General remarks .....	82
5.2 Random numbers .....	83
5.3 Questionnaires .....	87
5.4 Frames .....	97
5.5 Inadequate use of frames .....	102
5.6 Instructions .....	104
6. LISTING ERRORS .....	108
6.1 Introduction .....	108
6.2 Checking the quality of listings: Type I design .....	109
6.3 Estimation procedure .....	111
6.4 Presentation of results .....	112
6.5 Some comments .....	115
6.6 Type II design .....	116

6.7 Some illustrations .....	122
6.8 Summary of experiences .....	132
6.9 Problems in checking the quality of listing .....	135
6.10 Measures for improving the quality of listing .....	140
7. MISSING DATA .....	145
7.1 The problem .....	145
7.2 Consequences of missing data .....	146
7.3 The Hansen and Hurwitz technique .....	151
7.4 The Politz and Simmons technique .....	159
7.5 Other contributions .....	165
7.6 Success of the field work .....	168
7.7 Refusals .....	170
7.8 Post hoc adjustments for biases due to missing data ..	173
8. THE RESPONDENT .....	181
8.1 Introductory remarks .....	181
8.2 Intellectual background .....	181
8.3 Social background .....	184
8.4 Emotional background .....	189
8.5 Memory errors .....	191
8.6 Length of the period of reference .....	198
8.7 Continued observation .....	203
9. THE RESPONDENT (CONTINUED) .....	207
9.1 End effect .....	207
9.2 Open and closed reference periods .....	210
9.3 Location in time of the reference period .....	214
9.4 Conditioning .....	217
9.5 Variance considerations .....	220
9.6 Integration of errors .....	225
9.7 Checking the effects of errors due to respondents ..	234
9.8 Measures for improving the quality of the response ..	238

10. THE ENUMERATOR	242
10.1 Reasons for using enumerators .....	242
10.2 Some illustrations of enumerator effect .....	243
10.3 Definition of enumerator effect .....	245
10.4 Reasons for studying enumerator effect .....	248
10.5 Measuring enumerator effect .....	249
10.6 Some empirical studies of enumerator effect .....	252
10.7 General theory .....	255
10.8 Interpenetrating or replicated subsamples .....	262
10.9 Census applications .....	269
10.10 Some comments .....	278
10.11 Checking the quality of data collected by enumerators	282
10.12 Methods of improving the enumerators' work .....	284
11. SOME PROBLEMS OF QUALITY CHECKING	288
11.1 Reinterviewing .....	288
11.2 Timing check surveys .....	290
11.3 Advance knowledge of information being checked ..	294
11.4 Origin of errors .....	296
11.5 Joint effect of listing and response errors .....	297
11.6 Report of the check .....	300
12. CHECKING THE QUALITY OF PROCESSING	303
12.1 Introductory remarks .....	303
12.2 Aims of quality checking .....	305
12.3 Post hoc checking .....	305
12.4 Process checking .....	311
12.5 An illustration .....	319
12.6 Deliberate introduction of errors .....	322
12.7 Rational planning of data processing .....	323

13. ERRORS AND BIASES IN YIELD STATISTICS	331
13.1 Introduction .....	331
13.2 Selection of fields .....	334
13.3 Border bias .....	337
13.4 Location of plots inside fields .....	341
13.5 Plot size .....	342
13.6 Biases due to plot shape .....	347
13.7 Yield surveys of wheat planted in rows .....	348
13.8 Missing crop .....	350
13.9 Date of cutting .....	352
13.10 Yield surveys in small fields .....	353
13.11 Cutting procedure .....	353
13.12 Losses .....	355
13.13 Biases arising in the estimation procedure .....	356
13.14 Studying the quality of yield data by means of multiple cuts .....	359
13.15 Conclusion .....	365
14. CONCLUDING REMARKS	366
14.1 Role and importance of quality studies .....	366
14.2 Rational survey design .....	367
14.3 Pilot surveys .....	369
REFERENCES	372
AUTHORS' INDEX	389
SUBJECT INDEX	392

## 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

$$X = \sum_i^N x_i \quad (1.1)$$

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

$$z_i = x_i + d_i \quad (1.4)$$

and

$$\sum_i^N z_i = \sum_i^N x_i + \sum_i^N d_i$$

or

$$Z = X + D \quad (1.5)$$

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.

预览已结束，完整报告链接和二维码如下：

[https://www.yunbaogao.cn/report/index/report?reportId=5\\_22455](https://www.yunbaogao.cn/report/index/report?reportId=5_22455)

