UNITED NATIONS STATISTICAL COMMISSION
and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS
STATISTICAL STANDARDS AND STUDIES -- No. 2

AUTOMATIC EDITING OF INDIVIDUAL STATISTICAL OBSERVATIONS

UNITED NATIONS
New York, 1963

ST/CES/2

UNITED NATIONS PUBLICATION
Sales Number: 64.II.E/Mim.4

Copies of this document may be obtained from the Sales Section, European Office of the United Nations, Palais des Nations, Geneva, Switzerland, at the price of $0.76 (US), or may be ordered through the Distributors for United Nations publications in local currencies.

Preface

At its fourth plenary session in June 1956, the Conference o f European Statisticians established a Working Group on Electronic Data-Processing (originally Working Group on Electronic Data-Processing Machines) to exchange information on the experience of national statistical offices in using EDP for statistical purposes. The Working Group has met three times, in January 1957, in April 1961 (in Rome) and in December 1962. At its second meeting, the Working Group recommended to the Conference *inter alia* that it should meet again from time to time for detailed discussions of specific problems in the field of EDP and that such a meeting should be held in 1961/62 to discuss problems of input and/or editing and correcting data by EDP. At its ninth plenary session in July 1961, the Conference agreed that the next meeting of the Working Group should be held In the autumn of 1962 and should be devoted primarily to the discussion of problems of automatic editing and correcting of statistical data by EDP.

In making the arrangements for this meeting the Secretariat secured the services of an expert consultant, Mr. Svein Nordbotten of the Norwegian Central Bureau of Statistics to prepare a discussion paper on the subject of automatic editing and correcting of data. This paper was circulated with the reference number Conf.Eur.Stats/WG.9/37. The working Group "expressed appreciation of the value of the paper which drew attention to a very important practical problem in using EDP for statistical purposes, outlined a promising approach to the development of a general theory of the subject gave a most useful survey of different methods of automatically controlling and correcting statistical data, and put forward valuable suggestions concerning further work". The Working Group considered the paper in detail and the participants made a number of comments and suggestions on it. The Group recommended that the paper be revised and developed in the light of these comments and re-issued, The Conference of European Statisticians subsequently endorsed this recommendation and agreed that the revised paper should be issued in the published series of methodological studies of general interest "Statistical Standards and Studies".

While this paper has the general approval of the Working Group, it is not an agreed statement of the Group, but is issued in the name of the consultant and the secretariat.

GE.63-16648

AUTOMATIC EDITING OF INDIVIDUAL STATISTICAL OBSERVATIONS

by Mr. Svein Nordbotten
Central Bureau of Statistics, Norway
(Consultant to the Secretariat of the
Conference of European Statisticians)

Table of Contents
Paragraphs

Annexes

AUTOMATIC EDITING OF INDIVIDUAL STATISTICAL OBSERVATIONS

1. INTRODUCTION[*]

1. Electronic computers or data processing machines have for some years been used at several stages of statistical processing. The fields of application range from sorting, tallying, cumulating and editing to advanced mathematica1 and statistical computation. The aim of the present paper is to initiate a discussion about the use of electronic computers particularly for the control and correcting of statistical observations.

2. Objections to the use of electronic computers for editing statistical data have been made on two different and independent grounds. Some argue that an electronic computer and the experts working with it could be more efficiently employed on other tasks such as numerical computations. Others argue that editing statistical data is a skill, which requires the attention of experts, and cannot be approached automatically by means of methods derived from the theory of mathematical statistics.

3. It is hoped that this paper will help to convince statisticians that automatic computers may be both efficiently and successfully used in the editing of statistical data. This view was also recently expressed by Frank Yates of the Rothamsted Experimental Station in England, who stated: [1]

"As yet also, little has been done in research statistics on the general problem of the preliminary editing of data before analysis. This is a job, which is vitally important, even when the amount of data that has to be handled is quite modest. It is a job for which computers are theoretically eminently suited since once appropriately instructed a computer will perform any required tests on each item of data as it is read in, and can draw attention to anomalies, reject suspicious items or even in some cases make appropriate corrections. An excellent example of what is required is provided by the provision for error correction of time series data in the powerful general program for the analysis of time series (BOMM) described by Sir Edward Bullard at a recent meeting of the Society.

Here again is a task for statistical programmers in the immediate future. Provision for sophisticated data editing should, for example, form part of any good general survey program."

[*] The author is grateful to the Secretariat for assistance during the preparation of this paper and to the members of the Working Group on Electronic Data Processing for constructive comments received during the discussion of the paper at the Third Session of the Working Group.

[1] F. Yates: *Computers in research - promise and performance.* The Computer Journal, Jan. 1962, Vol. IV, No.4, pp. 273.279.

4. A discussion of this subject requires a definition of several concepts and a delimitation of the process denoted as editing. The second and third parts of the paper are devoted to these problems.

5. As mentioned above, electronic computers have already been used for some time for editing statistical data and the different procedures are examined in the fourth part of this paper. The survey is probably not exhaustive and it is hoped that it may be completed during future work in this field.

6. Insight into the problems of editing and the relative merits of different procedures can only be gained through systematic, empirical research. In the fifth part of the present paper a system for such research 1n this field is proposed. The aim is to be able to select methods possessing satisfactory general features. The research proposed in this paper is based on the construction of simulated observations with specified characteristics by means of Monte Carlo techniques instead of expensive observations, which require extensive as well as time-consuming inspection to satisfy our needs in research.

7. When an electronic computer is available, it is reasonable to use it in empirical research for simulating completely automatic as well as semi-automatic control and correcting procedures. The research system proposed has been programmed in Norway and the scheme is described.

8. An outline for future work and proposals for studies within the research system indicated are presented in the sixth part of the paper.

9. The author of this paper has on several occasions argued against the present procedures. They seem to be arbitrary and even inconsistent in some applications and therefore a theory of editing is needed. Automatic computers are very convenient for the intensive computations involved in using methods derived from such a theory, which are prohibitive if handled manually because they are too time-consuming.

2. PREPARATION OF STATISTICS

10. Editing statistical data is a part of the process of preparing statistics and it will be useful for further discussion to review the whole process. Such a review will of course be a simplification and idealization of the real process, but will help us in considering the editing as a part of a large process.

11. By preparation of statistics we mean any activity aiming at the provision of numerical information about masses of units according to a specified plan. This process can be divided into a number of operations characterized by their function as listed below: [2,3]

[2] P-J. Bjerve and S. Nordbotten: *Automatisering i statistikkproduksjonen*. Statistiske Meldinger, No. 6, Oslo 1956, pp. 1-17.

[3] M.J. Mendelson: *Data Processing Operations*, Handbook of Automation, Computation and Control, Vol. 2, John Wiley & Sons, 1959, Ch. 3, pp. 1-15.

(a) *Identification of the statistilca1 mass.*
(b) *Data collection.*
(c) *Data conversion.*
(d) *Data transcription.*
(e) *Data editing.*
(f) *Data arrangement.*
(g) *Aggregation of statistics.*
(h) *Presentation of statistics.*

The order in which the operations are listed is not necessarily the order in which they are actually performed. Data editing and transcription may for example often be interchanged.

12. Identification of the statistical mass is the practical counterpart to the theoretical definition of the mass, and is the operation needed for building up an identification register of all statistical units of the mass.

13. The operation necessary to obtain the data is called data collection, and this is performed in several ways as requests sent by mail to the respondents, by telephone calls, by personal interviews or direct observation, etc.

14. The data obtained are often in a form, which is not appropriate for modern processing, and need to be converted to another "language". Answers about profession need for example to be converted from verbal description to a numerical classification code. Usually these converted data must also be transcribed from the original questionnaires to other media such as punched cards, magnetic tapes, etc.

15. The above operations represent also sources of errors. A special editing operation is usually included in the process to detect and correct errors in the data. This operation may be performed at several places in the process depending partly on the type of the statistics and partly on the processing techniques applied. If manually performed, it may be done immediately after the data collection.

16. The data arrangement, aggregation and presentation of statistics lead to the statistical tables and their structures which are the final part of the process for preparation of statistics.

17. This final product may be characterized in several ways by:

(a) *Degree of information.*
(b) *Quality of information*
(c) *Currentness of information*

18. The degree of information describes for example the number and size of the tables produced, the quality of information expresses the overall accuracy of the statistics while the currentness is an inverse measure of the time between the observation and the completion of the statistics.

19. Improvement of any of these three aspects of the final statistics requires increased efforts and they therefore compete for statistical resources t0o do the work. None of the operations may, however, be improved by a change in processing techniques without a corresponding Increase in cost. The change from conventional equipment to electronic computers is perhaps one example of such a change which maw improves all aspects of the statistics if adequately utilized.

20. The improvement of statistical quality therefore should not be regarded as an independent aim, but should be considered together with the degree and currentness of information, and subject to overall cost. As the degree of information, i.e. the form of the tables, is usually determined first, improvement of quality frequently leads to less current statistics.

21. When an overall evaluation may indicate that improvement of the quality is needed, this does not necessarily mean that increased editing is the correct solution. There are types of errors, such as the coverage error due to incomplete identification of the statistical mass, which cannot be detected by editing, and there are response errors occurring in the data-collecting operation, which can be reduced more cheaply by improving the data-collecting methods than by extending editing. As will be emphasized in the following section, there exist also methods other than editing methods for continuous control of the quality of the process.

22. The editing, and in particular the automatic editing, which is the main subject of this paper, is thus only one of many possib1e ways for preparing better statistics of which the use of methods of collection which give more accurate observations is the most obvious. The decision to extend the editing operation should therefore be made only when an overaI1 assessment indicates that this is the most efficient way of improving the statistics.

3. CONTROLLING AND CORRECTING STATISTICAL DATA

3.1 Some basic concepts

23. The ultimate end of the production of statistical data is to gain information about the state and deve1opmenkt of society. This need for information has its origin in the abstract, theoretical systems or models representing our simplified apprehension of the real society. The components o f these conceptua1 systems may conveniently be denoted as theoretical variables.

24. The real world is, however, much more complicated than a human being can conceive and the observable, true variables are not usually identical with the theoretical variab1es defined within an idealized model picture of society. Trygve Haavelmo has expressed the difference in this way: [4]

"We may express the difference by saying that the "true" variables (or time functions) represent our ideal as to accurate measurement of reality "as it is in fact", while the variables defined in a theory are the true measurements that we should make lf reality were actually In accordance with our theoretical model."

25. This implies that the true variables can be measured even though it may be very expensive to do so. The value of the true variable is an aggregate or a function of the true, individual values of the units which make up the statistical mass. The individual true value must therefore always be

[4] Trygve Haavelmo: *The Probability Approach in Econometrics*, Supplement to Econometrica, Vo1. 12, Chicago 1944.

defined. operationally. Morris Hansen has suggested three criteria for the definition of "true value": [5]

1. *The true value must be uniquely defined.*
2. *The true value must be defined in such manner that the purposes of the survey are met. For example, in a study of school children's intelligence, we would ordinarily not define the true value as the score assigned by the child's teacher on a given date although this might be perfectly satisfactory for some studies (if, for example, our purpose was to study intelligence as measured by the teacher's ratings)*
3. *Where it is possible to do so consisently with the first two criteria, the true value should be defined in terms of operations which can actually be carried through (even though it might be difficult or expensive to perform the operations).*

26. It may be very difficult if not prohibitively expensive to perform the operations involved in measuring the true variable. The reason may be that a certain defined operation requires skilled experts or very large resources which are not available, and therefore the statistician resorts to short cuts which introduce errors. There are number of errors which may occur at different stages of a statistical investigation, a list of which is given by W.E. Deming. [6] These errors add up to total individual measurement errors. The variables as really measured are called observational variables and their values are called observational values.

27. The difference between the aggregated observational values and the true values is the measurement error, which the statistician strives to minimize. If the observed variables are subjected to a control and correcting process, say editing, we shall call the result edited variables from which the tables are computed.

28. To facilitate further discussion we shall use the following symbol notation. A letter with a prime or a double prime represents an observed or edited variable, respectively. For example:

x = individual true varlab1e
x' = individual observed variab1e
x'' = individual edited variable

29. The basic variables consist of the above *elementary* variables and the *derived* variable. The latter occur when non-statistical information such as theoretical knowledge, hypothesis concerning the users' model, etc., exists. From the basic variables some *auxiliary* variables are also derived such as error and correction components.

30. Each statistical unit may be characterized by several elementary variables and we shall denote the number of different elementary variables as the dimension of the measurement. From the elementary variables for each unit, we may be able to compute several derived variables by means of non-statistical information and we shall call the number of derived variables the *degree of knowledge*.

31. Then the true value of the m^{th} elementary variable of the i^{th} unit is denoted by x_{im} and the true value of the l^{th} derived variable of the i^{th} unit is denoted by:

[5] Morris H. Hansen and others: *Response Errors in Surveys*, Journal of the American Statistical Association, Vo1. 46, pp. 147.190, June 1951

[6] W.E. Deming: *Some Theory of Sampling*, John Wiley and Sons, New York 1950, pp. 24-52.

$$y_{il} = f_l\,(x_{i1} \quad \ldots\ldots x_{iM}) \qquad\qquad (l = 1.....\ L)$$

the functions representing our knowledge. The necessary auxiliary variables are defined by:

$$e_{im} = x'_{im} - x_{im} \qquad\qquad (i = 1......N)$$

$$c_{im} = x''_{im} - x'_{im} \qquad\qquad (m = 1.....M)$$

and

$$b_{il} = y'_{il} - y_{il} \qquad\qquad (i = 1.......N)$$

$$d_{il} = y''_{il} - y'_{il} \qquad\qquad (l = 1......L)$$

How these auxiliary variables should be defined is of course a question of convenience and they might equally well have been defined as multiplicative variables. But this would have complicated some of the later reasoning and they are therefore considered here to be additive.

32. Our considerations also include parameters, the values of which we try to fix subject to an optimum criterion. The parameters are for example the number of units which we want to correct automatically, the control limits, the coefficients of the correcting model, etc.

3.2 Control and correction

33. Given the fact that any piece of statistical information is affected by errors, their effect should be controlled and evaluated. Users of statistical information are mainly interested in the deviation between the total values of observed and true variables, while the statistician is also interested analytically in the errors occurring at each stage of the process in order to gain experience of how to allocate resources efficiently. This difference in the users' and producers' approach is important since the users' view alone does not justify the editing procedures currently applied.

34. Statisticians have tried to improve quality by introducing built-in control systems in the production of statistical data. These control systems constitute either control on a sample basis or complete control of all observations. The aim is to ensure that suspect lots or units are eliminated and reconsidered.

35. To evaluate the different control systems and to allocate resources efficiently require experience, which can often be gained through an analytical investigation. Allocating resources in order to keep the total errors within certain limits also demands an overall allocation theory for the production of statistical data. This seems to be an important subject because the extent of the quality control and editing should be determined in the light of the possibility of detecting errors in other stages of the statistical work, for example already during the observation.[7]

[7] Svein Nordbotten: *On Errors and Optimal Allocation in a Census*, Skandinavisk Aktuarietidsskrift, Uppsala 1958.

36. The statistical acceptance control often applied in the production of statistical data aims at the most efficient way of controlling the possible error which may be committed in classifying a lot of units as acceptable or not. The conventional editing procedure may be regarded as the extreme case where the number of units within each considered 1ot is one.

37. The main argument for controlling each unit instead of a sample is that at the editing stage it is impossible to foresee all future tabulations and therefore a complete control is necessary in order to avoid trouble. Because the tabulations at least in principle should be consistent with one another, all control has to be made before any tabulation starts. However, it should not be forgotten that editing may in particular cases be limited to only a sample, leaving resources available for the formulation of better editing rules.

38. This raises the question of the stage in the process of production of statistical data at which editing should be performed. In addition to depending on technical conditions such as the type of equipment available, there is a question of efficiency. The later the editing is undertaken, the more errors will probably be eliminated from the result, but the editing procedure will also be more complicated. The design of an editing procedure raises at least three main problems, i.e. the problem of specifying a *quality standard*, the problem of specifying a *control method* and the problem of specifying a *correcting method*, including certain identification problems.

39. The problem of specifying quality standards is very difficult and often overlooked. At this stage we only state that the quality standard is a quantitative expression of the degree of accuracy with which information is needed, i.e. about the extent of an interval around the observed value, which with a high degree of confidence includes the true value. If this interval is specified as equal to zero, it means that no errors are permitted at all.

40. The basis for any control procedure is some general knowledge about the structure of the mass to be measured and the basis for any correcting is the possibility of acquiring new values which can be regarded as more useful than the original ones. In control, this previous experience is used in specifying an acceptance zone.

41. There are two general types of such *a priori* knowledge which can be used in editing, i.e. *theoretical* and *empirical* knowledge. Theoretical knowledge is exact because we have ourselves imposed the definitions. The value of the variab1e *age* must for example equal the difference between the year of record and the year of birth. Theoretical knowledge gives us information which enables us to reject a set of values as not true, but it does not, except for some uninteresting cases, give enough in formation to indicate that a single value is untrue. These rejected values must, however, as stated by the British General Register Office, always be corrected.[8]

42. Theoretical knowledge is of great value and is the backbone of editing processes, but it is empirical knowledge which is the flesh and blood and which also gives rise to most of the problems. Empirical knowledge is utilized when the users set up their models, for instance when they are reasoning within a scheme based on the casual relationship between the input into a production process of raw materials, labour and capital equipment, and the output of products. Even though the relationship is hypothetical and is going to be tested by the information supplied by the statistician, it is both relevant and advantageous to take into account this vague knowledge in

[8] United Kingdom: *Report on Electronic Data Processing*, 1962 (Conf.Eur.Stats/WG.9/35/Add. 14).

editing to obtain more exact information which will give a more conclusive test. This is a fact which probably has been overlooked in most editing procedures.

43. Another type of empirical knowledge is statistical knowledge gained through previous processing of similar data or a decision sample. Statistical knowledge may consist of observed frequencies, correlations between variables, etc., which are not explicitly expressed in users' models. In foreign trade statistics, it has for years been a usual basis for control to check whether the ratio between value and quantity for each consignment is within limits outside which experience indicates it rarely occurs. Empirical knowledge may also comprise data on the value of the same variable at another point of time, for example data reported for previous years, or obtained by later observations.

44. While theoretical knowledge gives a basis for an *absolute* statement about the quality of observation, empirical knowledge must result in some kind of a *probability* statement e.g. that it is almost impossible that a given observation shows a true value.

45. All such knowledge ought to be systematically used in constructing two or more zones in which the observations are classified according to their observed values.

46. In accordance with what has been said in the Swedish report on automatic editing,[9] we may for example consider the following *three* zones in one of which any unit is classified according to its observed value:

l. *Zone for manual inspection of the value*
2. *Zone for automatic correction of the value*
3. *Zone for acceptance of the value*

In the first zone defective observations which may have a large effect on the results and perhaps need to be replaced by new observations are classified. In the second zone those rejected observations which affect the accuracy of the results less and which may be of more use if corrected by an automatic procedure are classified while the rest of the observations are classified in the third zone.

47. A specific type of action is associated with each zone. As to the first zone, the somewhat dubious assumption that manual inspection always leads to the true values is often accepted. This means that the relative efficiency of automatic correcting methods may probably be understated.

48. Empirical knowledge, together with quality standards, determines the limits of the zones. This relationship is much more complex and will be discussed in detail later. Usually the limits are, however, fixed rather arbitrarily by making rough use of empirical knowledge.

49. It should also be noted that several variables are usually observed simultaneously and that it is the space which is common to all the acceptance zones which defines the overall acceptance zone for the statistical unit.

50. Correcting, as used here, comprises all those operations in the production of statistical data which are performed to obtain new values for rejected observations. Correcting methods therefore include re-observation as well as automatic adjustment of the rejected values. If the rejected

[9] See Conf.Eur.Stats/WG. 9/35/Add. 3

observations affect the final aggregates seriously, the only way to increase the extent of information about the population being studied is to make new observations. Automatic correcting methods will only help to exploit effectively the information provided by the data, which have already been collected. This basic fact should always be emphasized when discussing automatic correcting in order to avoid creating the impression that the use of electronic computers for this purpose can serve as a substitute for the collection of improved data through further efforts at measurement.

51. In the correcting operation, theoretical knowledge alone is seldom sufficient and has to be supplemented by empirical knowledge. If, for example age does not equal the difference between the two points in time which define its value, it will be impossible to determine on a theoretical basis whether age, the first point of time, the second, both points of time or all three va1ues are wrong. If, however, we know empirically that it is almost impossible for the two points of time to be incorrect, we can correct the age with a high degree of confidence by substituting for the rejected value the time difference. Therefore, it will be impossible to make any absolute statement that the corrected value is true unless the unit is re-observed under completely ideal conditions.

52. In the correcting operation, the knowledge, which has been required concerning the population being investigated, is used to identify the errors and replace the rejected values by corrected values. The identification may for example be performed by means of a list of the variables associated with varying degrees of confidence determined from experience while the corrected values are supplied by means of a system of functions estimated from a sample of observations assumed to belong to the same populations as the observations to be edited.

53. In conclusion a warning should be sounded against uncritical control and correction based on empirical knowledge. If the statistician in his editing utilizes uncritically knowledge from the user's hypothetical models which is no more than a pure hypothesis that needs to be tested, and accepts correcting rules which in some way or another favor this hypothesis, it is obvious that the conclusion reached is influenced by the existence of the hypothesis itself. This may be an explanation of the phenomena of highly spurious correlations between observed variables.

4. APPLIED AND PROPOSED METHODS OF AUTOMATIC EDITING

4.1 Automatic versus manual editing

54. Let us first consider conventional, manual editing. The editing staff is composed of statisticians or others preferably with special knowledge of the type of statistical units investigated. To recruit people for this type of work with the necessary qualifications will usually be difficult, and the result is often an inhomogeneous editing staff with varying experience and knowledge.

55. The editing staff usually follows working instructions e those experienced in setting up working instructions know that it is necessary to choose between one of two alternatives. The instructions have either to be worked out in general terms leaving it to the staff to use their intelligence to solve particular problems, or they have to be detailed and probably the staff will

never be able to learn and apply them fully. In both cases, the editing will be left to individuals and the result will not be uniform.

56. As in many other fields, the statistician is apt to choose the most convenient solution and editing instructions are therefore often lacking in precision. Typical examples are the following lines taken from instructions for manual editing:

> "…..*look for an acceptable ratio between* …………"
> "…..*these values must be very small* ………………"
> "…..*unreasonable values must not be accepted*……."

57. The interpretation of "*acceptable*" , "*very small*" , "*unreasonable values*" , etc. is left to the individual editor. The way in which instructions will be read by the staff will probably vary according to education, experience and intelligence. More serious is the fact that the same person after six hours' work will act differently and after a couple of months he will probably have gained so much routine training that the results of his judgment will be very different from what they were at the beginning.

58. As previously pointed out, a major principle in the production of statistical data is that a defined relation expressed in specified operations ought to exist between observations and processed results. Manual editing both violates this principle and also introduces a new source of error, which may give rise to a special editing error due to imprecise editing.[10]

59. An advantage of automatic editing is that it is possible to introduce a very detailed set of editing instructions, which are strictly followed. This is illustrated by a quotation from the US Bureau of Labor Statistics:[11]

> "*Some time ago a "credibility" routine was scheduled to be introduced and a check was made of its efficiency. A file covering two months of data was used. This file had been edited, processed and corrected by clerks and hence was presumably error-free. The clerks had detected 323 inconsistencies, which they had corrected. The proposed mechanical testing uncovered 199 additional inconsistencies in this presumably error-free file.*"

Automatic editing therefore gives far more consistent processing than manual editing does. In addition, automatic editing will be performed faster and in general is less expensive, and therefore allows the application of more advanced methods than manual editing. It should, however, be strongly emphasized that neither computers nor people are able to make good data out of bad observations.

60. On the other hand, the preparations for automatic editing require extensive programming efforts, which means that the editing system might be rather rigid. If, for example, a logical error has been committed in constructing the editing program, the computer will consistent1y repeat this error throughout, which will require expensive new processing. Automatic editing therefore involves very accurate planning and preparation and any values and situations, which may occur,

[10] Svein Nordbotten: *Measuring the error of editing the questionnaires in a census*. Journal of the American Statistical Association, Vo1.50, 1955, pp. 364-369.

[11] See Conf.Eur.Stats/WG.9/35/Add. 13.

must be taken into account before it is started. Because of this inflexibility, it is highly desirable that the methods of automatic editing be designed in as general a way as possible.

61. Automatic editing has not been applied for a long time and there may be many advantages as well as drawbacks which are not yet obvious. As recently as ten years ago, a representative of the US Bureau of the Census, which has been a pioneer in the field, said:[12]

> "*A related line of development is the use of editing by mechanical tabulation equipment. During the 1950 census, increased use was made of such methods, thus eliminating some processes, which have been traditional. For example, it has been customary to subject schedules to editing processes for internal consistency, making also certain adjustments when they were not consistent.*"

In another paper from the same Bureau about 10 years later, it is reported that the automatic editing and tabulation represented as much as 36.7 per cent of the computer time used for the 1960 census of population in the period July 1959 - December 1960.[13]

62. Even though these progress reports are encouraging, statisticians ought to be aware of the fact very few methods of automatic editing have yet been tried. The automatic editing is by no means a specific statistical problem, but a special case of the general problem of automatic control and correction of information. Linguists, for example, are engaged in activities in this field, and statisticians may perhaps profit by studying their methods and results. L.E, Thorelli has recently treated a problem analogous to that of statistical editing in a paper in which he assumes that the original information is converted by a "*primary machine*" whose operations are subject to error. He then discusses the features of a "*secondary machine*" which is able to detect and correct these errors. He also discusses in his paper[14] different methods of which the "list method" resembles the "code control-method" considered later in the present paper.

63. In discussing automatic editing, it is convenient to distinguish three different phases. In order of application, they are:

(a) the numerical specification of the editing criteria,
(b) the control, and
(c) the correction.

However, we shall treat these three phases in a slightly different order leaving till last the methods for numerical specification.

[12] Morris H. Hansen: *The Statistical Research Organization of the Bureau of the Census*. Technical Paper No. 7 TA - OEEC - 105, 1952.

[13] Joseph F. Daly and A. Ross Eckler: *Applications of Electronic Equipment to Statistica1 Data-Processing in the US Bureau of the Census*, Bulletin de l' Institut International de Statistique 33e Session - Paris.

[14] Lars E. Thorelli: *Automatic Correction of Errors in Text.* BIT (Nordisk Tidsskrift for informasjonsbehandling) Bind 2, Hefte No.1, Lund 1962, pp.45-52.

4.2 Automatic control

4.2.1 The code control method

64. Automatic control was being applied in punch card techniques already in the first quarter of this century. The development of the electronic statistical punch card machine in connection with the census of population in 1950 is, however, a benchmark for large-scale automatic control.[15] The well-developed selection and counting capability of this type of machine made it possible to group and count cards according to multi-digit codes.

65. This ability was utilized in what we shall designate as the code control method which has later also been applied to electronic computers. In a population census the code 0 and 1 in a certain card column may for example denote that the person recorded is a male or a female, respectively, and other codes in this column have no defined meaning. A more advanced example of this type occurs when considering combinations of codes. If the person considered is recorded as a daughter of the family, this may be represented by the code 4 in another column of the card. We are now in a position to establish a control on the basis of the requirement that a card with code 4 punched in this second column can only be accepted as correct both as to position in family and sex if a code 1 is punched in the column recording sex.

66. The code control method only involves a few arithmetlca1 operations and has been widely used in connection with the 1950 and 1960 censuses of population and also in processing annual statistics in different fields. A very synoptic presentation of code controls in a census of population is given in the Finnish report on editing by computers. [16]

67. The application of the code control method by means of electronic computers in an integrated process requires refinements to save storage space and processing time. One refinement which is particularly well suited for binary computers has been used in connection with the 1960 census of population in Finland [17] and in several statistical applications in Norway.

68. This code control technique, which is particularly useful when the code list is constructed on the mnemonic principle and therefore contains many "*holes*" , is based on the idea that instead of storing a list of all acceptable codes or combinations and esting whether the code of the current record is within the list, each acceptable code is marked with a binary "*one*" in a binary string with as many positions as there are possible code combinations between the lowest and the highest. If, for example, the considered code is a combination of two one-digit codes for which the lowest value is 14 and the highest 361 the method required storage space for the two values 14 and 36 as well as string of (36 - 13) = 23 binary positions.

69. The method requires that the code x' currently considered is first tested to ascertain whether it is within the range 13 < x' <=36. Then if the code has a value of say 31, it is tested to ascertain whether a binary one is present in the (31 - 13]) = 18th position of the binary string. If not, the code

[15] Anthony A. Berlinsky: *Recently Developed Census Machines*. OEEC Technical Paper No.35, US Bureau of the Census. 1951.

[16] Central Bureau of Statistics of Finland*: Report on Electronic Data Processing,* (Conf.Eur.Stats/ WG.9/35/Add. 12).

[17] R. Kivivuori: *A Method for Checking Numerical Codes Using the 1401*, BIT (Nordisk Tidsskrift for informasjonsbehandling) Bind I, Hefte No.1, Lund 1961, pp. 48-53.

is rejected. The method indicates that if continuous codes can be used, it will only be necessary to test that the unit under current consideration has a code within the range of the code list (see *Figure l*). A similar control technique, which seems to be useful when the number of combinations is large, but related, is reported by the British General Register Office. [18] A three dimensional observation is first given a classification character by two of the three codes. This character is used as a key to a set of acceptable code patterns.

70. The examples mentioned above refer to code control based on theoretical knowledge. The method is of course also applied on the basis of empirical knowledge. The problem is then how to assign probable codes or combinations to an acceptance zone or list. A typical example from population censuses is that certain occupation codes are very unlikely to occur in conjunction with one of the two sex codes. A woman is rarely captain on a ship, while on the other hand a man is rarely a nurse. Those and other unlikely code combinations are regarded as undefined, i.e. a binary zero is present 1n the corresponding position of the string and the computer will reject the code combination either for manual Inspection or automatic correction.

71. So far, we have only referred to codes but the code control method can also be efficiently applied in controlling quantitative data. The range of variables which have a quantitative value, say a person's age, can be divided into a convenient number of classes associated with a code number and the code control applied to these code numbers. A well known application of the code method on converted quantitative data is the control of a mother's reported number of children against the year of marriage. The number of children and the duration of the marriage in years may here for example be considered as codes directly.

72. The code control method has also been applied in economic statistics. In Norway, the data reported in the 1953 Census of manufactures were automatically controlled by means of the code method.[19] The control was based partly on theoretical and partly on empirical knowledge. An example of the first type was that a firm established in 1953 by definition could only be accepted with s zero value of production in 1952. 0ne of the controls of the second type, which is illustrated in *Figure 2*, was based on an automatic inspection of the relation between value of production and employment classified by size groups. Both these variables were broken up into a scheme of classes associated with codes and the different combinations of the two codes were assigned either to the acceptance zone or to the rejection zone.

4.2.2 The ratio control method

73. The application mentioned above was based on the assumption that the value of production per man varies from establishment to establishment within an interval the limits of which can be fixed on the basis of previous experience. As modern electronic computers have a much larger capacity and speed in performing arithmetical operations than the punch card machines used at the beginning of 1950's a method previously used in manual editing which we shall call the *ratio control method* was adopted.

[18] United Kingdom: *Report on Electronic Data Processing*, 1962 (Conf.Eur.Stats/WG.9/35/ Add. 14)

[19] Svein Nordbotten: *Kontrollmetoder nyttet under bearbeidingen av bedriftstellingen 1953*, Monthly Bulletin of Statistics, No.12, Oslo 1955, pp, 333-339

74. Whereas the application of the code control method to quantitative data involves the conversion of the data to a code basis, the ratio control method calculates the ratio between two values and uses this derived variable as a control variable for which the acceptance zone is fixed. Acceptance of a ratio $R = y'/x'$ thus depends on whether the condition:

$$b_e < R <= b_u$$

is satisfied, b_e and b_u being the lower and upper limit of the acceptance zone respectively. With an electronic computer this is a much faster control method, though it should be noted that it involves the use of a strictly linear acceptance zone while in the code control method we are free to fix non-linear acceptance zones.

75. The ratio method has been used by many statistical offices especially in connection with foreign trade statistics and censuses of manufactures. In a work by the author of his paper, a description of the application of the ratio method in the Norwegian Central Bureau of Statistics is given.[20] Similar applications are reported by the Federal Republic of Germany,[21] the Netherlands,[22] and the United States. The description given for the foreign trade statistics in the above-mentioned paper, comprises both code control and ratio contro1. As to the latter, the ratio between the value and quantity of each repotted consignment is calculated and controlled against lower and upper limits of the acceptance zone or tolerance interval of the ratio. This acceptance zone is fixed so that any unit price outside its limits must be regarded as unlikely and is rejected for re-inspection.

76. An interesting extension of the ratio method can be obtained if the numerator or/and the denominator of the ratio is tested in addition against separate limits. The overall acceptance zone then takes the form of a polyhedron instead of a sector.

77. The use of the ratio method has been reported in several national papers on automatic data processing prepared for ECE meetings on automatic data processing. One extension of the ratio control method reported by the US Bureau of Labour Statistics is control of the ratios of the current and preceding month as a second stage ratio. [23] The use of historical information is often called *historical checks*, which seem to be very powerful. In connection with the automatic editing of the 1959 Annual Survey of Manufactures in the United States the use of historical information in ratio controls by means of a register system from the 1958 census, was particularly emphasized as valuable.[24] The British Board of Trade has also found that historical editing is generally worthwhile in spite of the extra work involved. [25]

78. Another interesting development of the method is used in the US Bureau of the Census.[26] In this version of the method records are assigned to one of three zones, i.e. the acceptance zone, the zone of non-classified records or the rejection zone. As in the simple ratio control method, a record

[20] Svein Nordbotten*: Statistical Data Processing in The Central Bureau of Statistics of Norway*. Bulletin de la Institut International de Statistique 33e Session, Paris 1961.

[21] The Federal Statistical Office of Germany*: Report on Electronic Data Processing*, May 1962.(Conf.Eur.Stats/WG.9/35/Add.2).

[22] The Netherlands Central Bureau of Statistics: *Report on Electronic Data Processing* May 1962 (Conf. Eur.Stats/WG. 9/] 5/Add. 5).

[23] See Conf.Eur.Stats/WG.9/35/Add.13.

[24] US Bureau of the Census: *Processing the 1959 Annual Survey of Manufacturers*, Jan. 1961, Memorandum.

[25] United Kingdom: *Report on Electronic Data Processing*, 1962 Conf.Eur.Stats/WG.9/35/Add.14)

[26] US Bureau of the Census: *Specification for UNIVAC Processing of Foreign Trade Statistics*, 1953, Memorandum.

is accepted as to value and quantity, if the unit price is between the lower and upper limits, otherwise it is assigned to the zone of non-classified records. The figures relating to quantities and those relating to values are aggregated for all records as well as for the non-classified records separately. At the end of each commodity group the average price ratio of all records is calculated and this ratio is subjected to a second ratio control with narrower limits. If accepted, or if both the ratio of the value sum of non-classified records to the value sum of all records and the ratio of the quantity sum of non-classified to the quantity sum of all records satisfy separate ratio controls of all records in the non-classified zone are transferred to the acceptance zone. Otherwise, they are transferred to the rejection zone for either manual inspection or automatic correction.

4.2.3 The zero control and the functional control method

79. The third control method in this section is the *zero control method*. This method is based on theoretica1 knowledge of the type:

Value added by manufacture + cost of raw materials, etc. + contract work
- gross value of production = 0

The zero control method has been used as a technical control method in data processing for a long time, but it is now a1so applied as a method for controlling the individual reports. As may be imagined, most applications refer to accounting data.[27, 28]

80. Related to this zero control method is the *functional control method*. This method requires that the value of some arithmetical expression, say the sum of two or more variables, must lie between specified 1ower and upper limits in order to be accepted. The general criterion for accepting a record subject to this control is thus:

$$b_e < f\,(x_1', \ldots, x_n') <= b_u$$

where b_e and b_u are limits and $x_1', \ldots, x_n'$ the recorded values subject to contro1. The ratio and the zero control methods are in fact special cases of the functional method. Another important case appears when the function includes one variab1e only.

81. Not all variab1es in the function need be subject to editing. In the case of historical editing, the historical variab1es are assumed to represent true values. They may, however, be related to current variables in some way and their inclusion will therefore strengthen the control. This may be called the principle of conditional editing.

82. The following application of the *functional control method* has been studied in Norway.[29] In

[27] Inter-American Statistical Institute: *Compilation of External Trade Statistics of Latin America by Computer*, Paper prepared for the Meeting of the Expert Group on International Compilation of External Trade Statistics, Rome, Feb. - March 1962

[28] See Conf.Eur.Stats/WG.9/35/Add.13

[29] Svein Nordbotten: *Maskinell revisjon*, Central Bureau of Statistics, Oslo 1956, Memorandum.

econometrics, an exponential relationship between output and the input of labour, use of capital equipment etc. called the Cobb-Douglas function of production has been traced. This knowledge might be used in a control of the types for acceptance:

$$b_e < x' (n')^a * (c')^b <= b_u$$

where b_e and b_u are the lower and upper limits, x', n' and c' denote quantity produced, number of man-hours and measure of capital equipment utilized, respectively. The *a* and *b* are two exponents to be fixed by some means, for example by regression techniques as proposed by Frank Yates.[30] If n' is obtained from another source, and can be assumed to be correct, this would illustrate the use of a conditional control.

83. The functional control method allows particularly for taking care of the user model and the control functions may be regarded as a miniature picture of the average statistical unit.

4.2.4 The gross error control method

84. In a paper presented recently at a meeting of a Scandinavian committee for technical co-operation, a control method not requiring pre-fixed limits was proposed. [31] The *gross error control method* is based on the paper by W.J. Dixon about ratios involving extreme values[32], and assumes that we are dealing with a normal distribution. This may be a serious objection to the method.

85. In contrast to the previous methods, data or ratios from more than one unit are dealt with at the same time. For example, we may deal with successive lots of three units. The three values are first listed in order of size, e.g. $x_1'<=x_2'<=x_3'$, and the following two expressions are calculated:

$$r_{10} = (x_3' - x_2')/(x_3' - x_1')$$

$$r_{01} = (x_2' - x_1')/(x_3' - x_1')$$

Subject to a normal distribution of x, the probability that x_3' is correct is *0.05* if r_{10} exceeds *0.97* in value. The same is true for x_1' if r_{01} exceeds *0.97*, which may be used as an upper limit for an acceptance zone.

86. The method must of course be modified in order to fit our requirements. Moving samples might be used so that two or three consecutive values could be rejected.

87. In the general discussion of editing we stated that editing could be regarded as an extreme case of statistical acceptance control. The main reason for not applying acceptance control based on sampling instead of 100 per cent editing, is that the statistical records may be intensively classified and a relatively small error in an uncontrolled record may influence the results for a particular sub-group seriously.

[30] Frank Yates: *Sampling Methods for Censuses and Surveys*, Charles Griffin and Co., Third Edition London 1960, pp~ 392-393

[31] Svein Nordbotten: *Notat om et simuleringssystem for vurdering av automatiske granskningsmetoder*, Central Bureau of Statistics, Oslo 1962, Memorandum.

[32] W.J. Dixon: *Ratios involving extreme values*, Annals of Math. Stat., Vo1. 22 (1952). pp. 68-78.

88. In special purpose statistical surveys the tabulations may be well known and the risk of using the *acceptance control method* may be fount to be negligible. However, this method can only be used if the variables used in establishing the control are normally distributed.

89. In contrast to the quality control applied for example to coding and punching operations, for which the main interest is to keep the risk α of rejecting a group which should be accepted as small as possible, the main aim of the acceptance control method is to ensure that groups of records do not include serious errors, i.e. we want to keep the risk β of accepting a group of observations which should be rejected as small as possible.

90. For each statistical group an acceptance zone for the standard deviation of the variable to be controlled is established by means of an upper limit b_u for the value of the standard deviation. Thus the acceptance zone is:

$$s <= b_u$$

Using the standard deviation as a control variable is reasonable, The limits of the ratio control may mean tbat the probability is only 0.01 that a correct ratio will be outside the limits. If, for example, the ratios are normally distributed, then:

$$s = (b_u - b_l)/6$$

will be approximately true.

91. As we are now dealing with sampling, the application of the acceptance control method requires also for each group the number of records to be sampled and controlled by the computer. This number can be computed when deciding that the risk of accepting a standard deviation larger than an alternative limit, say b_α, should be below a specified level β in addition to the requirement that the risk of rejecting a group with an acceptable standard deviation shou1d be less than α. If the levels of α and β and the ratio b_a / b_u between the two alternative limits do not vary from one group to another, the size of the sample, n, will also be the same and independent of the size of the statistical groups.

92. When starting the control, the computer needs the parameters n, b_u and $\chi^2_{\alpha,\, n-l}$. The last is the value of the chi-square statistic corresponding to the confidence level α and the number of freedoms, $n - l$. This value can be regarded as a transformation of the upper limit b_u. For each group, n records are drawn at random. The expression

$$k^2 = (n-1)*s^2/b_u^2$$

is computed and considered as our control variable. If this computed value exceeds χ^2, which gives the upper limit of our acceptance zone, we reject the whole group, otherwise all records are accepted with the risk that the s is not within its acceptance zone.

93. Obviously, the most efficient size of the groups depends on the frequency of rejections and larger statistical groups should probably be sub-divided into sub-groups, which are controlled separately.

94. A refinement of the acceptance control method is the sequential plan for acceptance control based on the ideas of Abraham Wald.[33] This type of acceptance control seems to be particularly well suited for application by means of electronic computers. As in the case of the preceding method, we need to specify b_u, b_a, α and β, while the sample size n will depend on the structure of the population and will never be larger than necessary to reach a decision. The method is therefore on average faster than the preceding one.

95. The control is commenced by computing the limits of three zones on the basis of b_a, b_u, α and β. These limits are functions of the sample size n and refer to the acceptance zone, the "continue" zone and the rejection zone for a statistic Z. For each group the computer draws a sample at random and computes the value

$$Z <= (x_i' - \bar{x}')^2$$

as long as Z is classified in the "continue" zone. As soon as Z is classified in either the acceptance or the rejection zone, the sampling is terminated and the whole group of records is accepted or rejected.

96. The standard deviation has here been proposed as a convenient control variable but in many control problems the range would perhaps be as appropriate.

97. The assumption of normal distribution is the main weakness of this method as the basic characteristics usually have skew distributions. This might be solved by using derived expressions of two or several characteristics as control variables.

4.2.6 Other control methods

98. An approach based on the computation of discriminant functions has been proposed by Robert Ferber and is outlined in the report by the US Federal Reserve Board.[34] Reports on the same items from two sources are paired and discriminant functions computed. These are used in water surveys to identify respondents with reporting problems.

99. John Tukey has recently devised a control method for rejecting "wild shots" in a set of data,[35] which he calls FUNOR and which lends itself to application by automatic means. *FUNOR* stands for Full Normal Rejection and is based as the name indicates on the assumption of normal distribution.

100. The method involves an ordering of all observations by size and an automatic inspection of the *1/3* of the observations with the smallest values and the 1/3 with the largest values for which the expression

$$z_i =(x_i' - \bar{x})/a_{i/n}$$

is computed, where $\bar{x}$ is the median and $a_{i/n}$ is the typical value for the i^{th} ordered value in a set of n observations from *8* normal distribution with a standard deviation equal to *1*. Then the median,

[33] Abraham Wald: *Sequential Analysis*. John Wiley and Sons, N.Y. 1947, pp. 125-133.

[34] See Conf. Eur. Stats/WG, 9/35/Add. 13.

[35] John W. Tukey: *The future of data analysis*, Annals of Oath. Stat. Vol.33, No.1, March 1962, pp. 1-68.

z, of the approximately *2/3 n* observations for which z_i is calculated, is deduced.

101. Each value z_i is controlled against an acceptance zone determined by:

$$|x_i{}' - x| >= A*\bar{z} \text{ and } zi >= B*\bar{z}$$

where *A* and *B* have to be specified by some adequate method. This control method can be regarded as another special case of the functional control method.

4.3 Automatic correcting

102. While many statistical offices have reported experience in the application of automatic control, few seem to have applied automatic correcting methods. This is, of course, due to the fact that the correcting phase of the editing is more complicated. The statistical bureau with the most extensive experience in this field, the US Bureau of the Census, emphasizes that corrections which may have an important effect on the aggregates should not be done automatically, but through further efforts at measurement. In the Bureau of the Census the reasons for employing automatic correcting are often to avoid unnecessary programming or using printing space for small "unknown" categories. The number of correcting methods tried is therefore rather small, even though several of those proposed have not yet been examined. The following quotation indicates, however, that the problem of automatic correcting, is regarded with great interest in statistical offices.[36]

> *"To solve the automatic correcting is in our view one of the most important technical problems for the production of statistics if you want to use the EDP-equipment efficient1y."*

103. In discussing automatic correcting methods we should, however, keep in mind that they do not contribute to the extent of information. Automatic correcting is a method of exploiting the existing information as fully as possible for the benefit of the user of statistics. The automatic correcting problem seems to be very similar to the non-response problem.

104. The terminology in this field is, of course, also very varied. Besides the distinction between control and correction, it has also been usual to separate the problem in two c1asses, according to incorrect values and missing values, respectively. From the author's point of view there is very little to be gained by this distinction as long as the problem is handled automatically. Both require a method for estimating a better or more probable value and we therefore denote all methods as *correcting methods*. In the US Bureau of the Census the corrected values are called imputed values, allocations or assignments, and are made when information "*was lacking on the schedules or when certain information reported was inconsistent.*"[37]

105. There is, however, another aspect of the correcting phase which calls for attention. Frequently, the control method rejects a record because the value of a function is incorrect or unlikely, but does not indicate which of the variables within it ought to be corrected. Therefore,

[36] Central Bureau of Statistics of Sweden: *Report on Electronic Data Processing*, April 1962 (Conf.Eur.Stats/WG.9/35/Add.3).

[37] US Bureau of the Census: *General Population Characteristics*, 1960 Census of Population, 1961, pp. XVII-XVIII.

some correcting methods will comprise both *identification* and a *correction* stage. When applying, for example, the ratio control method which rejects a record because the derived ratio between the value and quantity is outside the acceptance zone, it has to be decided whether value or quantity is incorrect or perhaps both.

4.3.1 The cold deck correcting method

106. The cold deck method is one introduced already by the US Bureau of the Census in connection with the 1940 Census of Population as a method for imputing missing values.[38] The idea of the old method is presented in this paper as a method for correcting. The *cold deck correcting method is* based on a cross-classification scheme designed especially for control purposes. The classification scheme is constructed in such a way that from a correcting point of view there are as great differences as possible between records belonging to different classes whereas the differences between two records within the same class are insignificant. For each cell in this scheme at least one representative record is stored in the computer.

107. When a record is rejected because one or more of its values are outside the acceptance zones, the record is classified with reference to the rest of the variables. At least one representative record should now be stored in the computer, with the same classifications giving representative values to replace those rejected In the considered record. If there are several representative records in the same classification, one is selected either systematically or at random.

108. If, for example, age, sex and marital status are variables recorded in a population census, the cold deck method will require representative records for each acceptable combination of sex and marital status to give a correcting value of age. When the control has rejected the age in a record as probably wrong, the record is then classified according to sex and marital status. There may be several representative records with this particular classification by sex and status giving different representative ages. When two or more records exist, the computer selects one in accordance with a specified rule. The age of the selected representative record is used as the corrected age in the considered record.

109. A typical selection rule is given by the following example. The information about sex may be missing. The record is classified in one editing class for which "male" and "female" occur with equal frequency. Therefore, two representative records are established1 one with the value "male", the other with the value "female". Each odd missing-value-record is corrected by means of the first representative record, and each even record by the second.[39]

110. When, as In the above examples, we are considering only the basic, elementary variables, the cold deck correcting method does not create any identification problem. Such a prob1em does arise, however, when applying, for example, a ratio control in foreign trade statistics. In this application neither value nor quantity is rejected, but the ratio is if it falls outside the acceptance zone, and an identification method is needed to decide whether value or quantity or perhaps both should be corrected.

[38] Howard G. Brunsman: *Processing and editing the data from the 1960 Census of Population*, US Bureau of the Census, 1960.

[39] See Conf .Eur.Stats/WG. 9/35/Add. 13.

111. The simplest method - which has been tried on an experimental basis in Norway - is to assign different degrees of confidence to the variab1es involved. This principle implies that if one variable is considered as incorrect, all variables with a lower degree of confidence are also considered incorrect. In the case of foreign trade statistics in which three characteristics, i. e. the commodity code, the value and the quantity, are subjected to code and ratio controls, the confidence degree of the characteristics may be - in decreasing order - code, value and quantity. If the reported value is outside its tolerance interval both quantity and value are assumed to be incorrect. On the other hand, if only the ratio is rejected, the quantity is considered to be incorrect. For each commodity code a cold deck consisting of two sets of representative records are needed. The first set gives representative values for the commodity groups for both value and quantity and may only contain one single record, which of course will only give rough estimates. This set of records is used when the value is rejected. The second set gives representative quantities for the different classes of values. This set must at least contain as many representative records as there are value classes, and is used to get a corrected quantity when only the ratio is rejected.

112. A similar identification principle is now applied in the automatic processing of the foreign trade statistics of the United States because they have found 1n numerous studies that value is one of the most reliable data items recorded.[40] The British Board of Trade also reports that in enquiries in which data on both quantities and values are collected, the quantity returned is much more likely to be the cause of error than the value.[41]

4.3.2 The hot deck correcting method

113. The method described in the preceding section requires a deck of representative records based on experience, which are stored in the computer. It may be very difficult because of storage capacity etc., to take into account all characterlst1cs such as, for example, variations in the material from one district to another. Another limitation of the cold deck method is that it does not make any use of the current data.

114. Howard Brunsman describes in his paper another imputation method, which takes into account these objections. The method is called the *hot deck method* and was applied in connection with the 1960 Census of Population in the United States. As with the cold deck method, the hot deck method is presented here as a general correcting method. The idea behind the hot deck is that the representative records are currently adjusted by the values of each accepted record.

115. To exemplify the method we return to the population census with sex, marital status and age reported. Before the correcting process can commence, the computer must be supplied with an initial cold deck of representative records covering all acceptable combinations of sex and marital status. Both rejected and accepted records are classified according to this method, the rejected records to be corrected, and the accepted records for adjusting the representative records with more up-to-date values. The age in an accepted record will for example be used as an adjusted value of age in the representative record with the same sex and marital status combination. If the age of the next record with the same sex and marital status combination is rejected, it will be replaced by the age of the preceding accepted record with the same combination. (see *Figure 3*).

[40] F.A. Scharff, N. Swersky and E.L. Wendt: *Some implications of computer processing of economic censuses and surveys*, US Bureau of the Census, 1961, Stenciled.

[41] United Kingdom: *Report on Electronic Data Processing 1962* (Conf.Eur.Stats/WG.9/ 3 5/Add. 14)

116. In connection with the editing of the *25%* sample taken at the time of the US 1960 Censuses of Population and Housing, much more extensive systems were applied with editing classification schemes with up to 1,500 cells, each, stored and adjusted in the computer during editing.[42] In this connection, very intensive use of the method was also made in substituting information about persons for whom no record existed at all, accounting for about *0. 5%* of all persons covered by the census.

117. The hot deck correcting method seems to be a very flexible means of taking care of the trends in the structure of the population. It should be noted that, considered separately, this method requires that the whole data mass - including accepted records - be run through the computer and is therefore well suited for integration with the control process.

118. It should also be mentioned that the applications in the United States include a procedure for recording in a diary the number of corrections made which is later used in indicating the quality of the results and for rejecting the whole statistical group if the number of corrections exceeds a given upper limit.

4.3.3 The Monte Carlo correcting method

119. Gunnar Andreasson has proposed the use of the Monte Carlo technique for correcting rejected values.[43] His *Monte Carlo correcting method* assumes random drawings on the basis of the cumulative distributions of the true values of the considered variables. The distribution is built up on accepted values which are supposed to represent the true values during the correcting phase. By a random number generator a random function value is obtained and through the cumulated distribution function the corresponding argument or variable value is derived and used as an estimate.

120. In the example from a census of population, correction of age implies that cumulative distribution tables of age ought to be built up for each combination of sex and marital status during the control and then stored in the computer before correcting commences. Each record with a reported age which is not accepted has first to be classified by sex and marital status to select the relevant distribution. Then a random number has to be generated and used to obtain a new age value by means of the distribution. This value is then used as a corrected value in the rejected record. As the distributions are cumulated from accepted records, we are sure that the corrected values will always be acceptable.

121. Compared with the cold deck correcting method this method preserves the distributive characteristics of the accepted records whole the cold deck method tends to give an artificial concentration around the values of the representative records. If the records are in random order, the difference between the Monte Carlo and hot deck methods seems only to be on the surface but if there is a hidden ordering or stratification of the records f which is the ordinary case, this will make the hot deck method superior. In contrast to the bob deck correcting method, it also requires separate runs through the computer for the control and correcting operations.

[42] US Bureau of the Census: *Editing and weighting of the sample population and housing data*, February 1960, Memorandum.

[43] Gunnar Andreasson: *7070 program för simulering av automatisk rättning med Monte Carlo-teknikk vid statistisk tabulering i datamaskin*, Central Bureau of Statistics, Stockholm 1961.

122. There is a special situation, in which the last statement is Invalid. If the distribution can be assumed *a priori* such as the distribution of the 1ast digit of the birth year which may be rectangular, the correcting may proceed in parallel with the contro1, for example by systematic drawing from this distribution.[44]

4.3.4 The functional correcting method

123. We shall call the general correcting method mentioned here the *functional correcting method*. This method is the counterpart to the functional control method. The basis for this method is a set of numerically specified relations stored in the computer:

$$f_i\,(x_1'',\ldots..,\,x_M'') = 0 \qquad\qquad (i = 1,\ldots\ldots.,\,M)$$

124. Some of the variables included may very well represent values from "outside" such as edited values from previous investigations. This type of conditional correcting is going to be used to a large extent in processing the 1963 Census of Manufactures in the United States, but was applied already in the corresponding census in 1954 in which conditional automatic correcting was performed by means of payroll information obtained from the Bureau of Old Age and Survivors Insurance.[45, 46]

125. The relations may be ordered by degree of confidence, i.e. if only one value is rejected the first function is used to estimate the corrected value on the basis of the accepted values in the record. When two values are rejected, the first and second functions are applied, and so on.

126. Let us consider foreign trade statistics, the records of which are characterized by code1 value and quantity, and assume that code and ratio controls are applied. Based on the accepted records or other information the normal or average values of value, x_l, and quantity, x_2, are applied for each code group.

127. When the ratio x_1'/x_2' is rejected the identification of the incorrect value may be done by a fixed confidence degree connected to each variable. Another method proposed is to regard the variable with the largest relative deviation from its mean, *m (x)*, as the incorrect one, Thus, if

$$a*(x_1'-m\,(x_l))/m\,(x_1)\;\; -b*(x_2'-m\,(x_2))/m\,(x_2) >= 0$$

where *a* and *b* are weights both of which first can be supposed to be equal to *1*, this means that x_1' is wrong. On the other hand, if the expression is negative, x_2' is considered incorrect. Proposed modifications are to use the standard deviation as the denominator, and to determine an appropriate value for *a/b*.

128. When the wrong value or values are identified, the following correcting functions are typical for the functional method:

[44] US Bureau of the Census: *1960 population sample-computer edits*, Sept. 1960, Memorandum

[45] US Bureau of the Census: *Notes on general plan for computer processing of the 1963 Census of Manufactures and Mineral Industries*, May 19G2, Memorandum.

[46] Maxwell R. Conklin and Owen C. Gretton: *Some experience with electronic computers in processing the 1954 Census of Manufactures*, Paper presented to the Annual Meeting of the American Statistical Association, 1957.

$$f_1\ (x_1'', x_2'') = x_1'' - x_2''{*}m\ (x_1)/m\ (x_2) = 0$$
$$f_2\ (x_1'', x_2'') = x_1'' + x_2'' - m\ (x_l)\ m\ (x_2) = 0$$

The reasoning behind this correcting model is that a corrected value taking into account the accepted value is represented by the first relation, while if both values are rejected the mean values will be good corrected values. In a record in which x_1' is identified as incorrect, the corrected value will thus be:

$$x_1'' = x_2'{*}m\ (x_1)/m\ (x_2)$$

If both x_1' and x_2' are found to be wrong, both functions are used to determine the corrected values. It can easily be seen that these values will be $x_1'' = m\ (x_l)$ *and* $x_2'' = m\ (x_2)$. A similar technique is illustrated in *Figure 4*.

129. As with all examples appearing in this paper, the above example is simplified. A similar, but much more complicated approach was, however, applied in the 1958 Census of Manufactures and later Annual Surveys in the United States.[47]

130. The first objection to this type of correction model will be that the standard deviation of the population will be too greatly disturbed. If the numerical values of the model are specified by some statistical estimation procedure, it is possible to add a random component to the corrected values by applying the Monte Carlo technique.

131. The second objection may be that the correcting method ought to take into account the observed values because they contain some information even though it is not as good as is desired. This can be accomplished by adjusting the corrected values by, for example, the deviation of the incorrect value from its mean.

132, Another obvious type of functional correcting is obtained if the corrected values are produced by regression equations. This was tried on an experimental basis in connection with the 1955 Annual Production Statistics in Norway with promising results. Tore Dalenius has also pointed out that this problem is related to that of missing values in sample surveys which was treated by S.F. Buck recently.[48]

4.3.5 Other correcting methods

133. In connection with the FUNOR control method, John Tukey also proposes a value modification or correcting method, which he calls FUNOM (Full Normal Modification). The method gives as an approximate overall result that deviations rejected in respect of the FUNOR

[47] US Bureau of the Census: *Industry Edit Specifications and 1958 Census Computer specifications*, July 1958, Memorandum.

[48] S.F. Buck*: A method of estimation of missing values in multivariate data suitable for use with electornic computer*, Journal of the Royal Statistical Society Series B, 1960, 00. 302-306.

contro1 are corrected up to the median value while other values - not rejected, but deviating more than $B_m * a_{i/n} * z$ where B_m is predetermined - are replaced by this value.

4.4 Methods for numerical specification of editing criteria

134. The application of the different control and correcting methods requires a means of specifying the editing criteria numerically. This operation will also imply consideration of quality requirements, which in this paper will be considered as predetermined. The methods for numerical specification may be group Into two main classes depending on whether the specification ls based on subjective judgment or statistical methods.

135. We shall look upon the problem as a *two-stage process*. The first stage will be an intensive analysis of a sample of units which is here assumed to give complete knowledge of the true values of the sample units, This sample, called the *decision sample*, represents our statistical knowledge about the population, and is the basis for deciding (subject to quality standard requirements) whether editing is necessary. If so t in the second stage of the process the *statistical* knowledge gives information necessary for the specification of control and correcting criteria implied by a particular method of editing. The decision sample gives us the additional information, which entitles us to argue that automatic correcting may increase the quality of the statistical results.

136. The decision sample does not need to be a random sample, but if any objective evaluation is needed of the editing method and the results it yields t a random sample is required.

137. The performance of an editing procedure must of course depend on the requirements of the statistical measurement, if the requirements are weak1 perhaps no editing is necessary at all, but if they are tight it may - for example in the case of ratio control - imply very narrow limits to ensure that most observations influenced by errors exceeding a certain magnitude are rejected for inspection. Tight requirements also have implications for the automatic correcting procedure, Applying the functional correcting method, tight requirements may demand that the correcting functions should include as many variables as possible to reduce the variance of the corrected value to the required minimum.

138. As most of the control and correcting applications rely on empirical knowledge, which implies that conclusions have to be reached on an inductive basis, the control as well as the correcting must be performed in a probability sense and it is also necessary to express the quality standard as a probability statement. The required quality standard has therefore to be of the type

$$P\ (|X''- X| > a) = p$$

i.e. the probability that the total error, $|X'' - X|$, is larger than a specified value, a, is equal to p which is usually set at a low value.

139. Any editing procedure ought to be designed in such a manner that the requirements as to quality are met. We shall limit our discussion to the case where the population means are the characteristics measured and any requirements will therefore be related to these characteristics. We shall never be able to say anything exact about the quality. Therefore any objective statement must be a probability statement and any decision must be subject to certain risks.

140. We first wish to testy on the basis of the statistics computed from the decision sample, the following set of hypotheses to decide whether any editing is necessary at all:

$$H_0\text{: } E\,m\,(e_m) = 0, \qquad \text{against } H_1\text{: } E\,m\,(e_m) \neq 0, \qquad (m = 1,\ldots, M)$$

$$H_0\text{: } E\,m\,(b_l) = 0, \qquad \text{against } H_l\text{: } E\,m\,(b_l) \neq 0, \qquad (l = 1,\ldots., L)$$

where *E* stands for the mathematical expectation of the estimates, *m* (e_m) and *m* (b_l) stand for the estimates of the means.

141. From the point of view of the producer of statistics, we want particularly to be on guard against the risk of rejecting H_0 when it is true, because we will then start unnecessary and expensive editing which may reduce instead of increasing the quality of the results. If we decide to take a risk, which we cannot avoid, of α%, we can establish the well-known criteria for rejecting H:

$$\left| \frac{m\,(e_m)}{\sqrt{\dfrac{(N-n)*m\,(e_m*e_m)}{N*n}}} \right| > t_{\alpha,n-1} \qquad (m = 1,\ldots\ldots, M)$$

$$\left| \frac{m\,(b_e)}{\sqrt{\dfrac{(N-n)*m\,(b_l*b_l)}{N*n}}} \right| > t_{\alpha,n-1} \qquad (l = 1,\ldots\ldots., L)$$

where the $t_{\alpha,n-1}$ is $\alpha\%$ value of a *t*-distributed variable with $(n-1)$ degrees of freedom, and $m\,(e_m*e_m)$ and $m\,(b_l*b_l)$ are the second order central moments of *e* and *b*.

142. Let us consider the test from the user's point of view. The user is not generally interested in whether the statistician edits or not, provided he can trust the results. The user will mainly be interested in avoiding the second type of error, i.e. that H_0 is accepted by the producer when an alternative H_1 is true. The user may, however, tolerate an error within seasonable limits, for example up to a fraction *k* of the measured value. As any decision must be made on a sample basis, the producer cannot guarantee anything with certainty, and to get results, the user will have to be willing to take a risk β that the error in the result may exceed $k*m\,(x')$. In other words, as has already been indicated above, the user's quality standard requirements have to be formulated in the following manner:

$$P[|\, E\, m\, (e_m)|\; > a_m\, (= k^*m\, (x_m'))] \;\; = \beta \qquad\qquad (m = 1,......, M)$$

and

$$P\, [|\, E\, m\, (b_l)|\; > a_l\, (= k^*m\, (y_l'))] \quad =\beta \qquad\qquad (l = 1,……., L)$$

The two risk levels, the hypothesis and its alternative, and the sample size are interrelated by operating characteristics curves, which determine the necessary sample sizes:[49]

$$n = n\, (\alpha, \beta, m\, (e) =0, a\, m\, (e^*e))$$

The moment term indicates that a guess is required as to the variance of the error-variable. As the sample size is predetermined in practice, the risk level α will usually be the term determined through the operating characteristics curve.

143. When an H_0 hypothesis is rejected, editing will be necessary and we shal1 have to design the editing methods in such a way that the quality standard requirements will be satisfied. After editing, the following criteria, called the *first set of criteria* must therefore not be vio1ated:

$$t_m = \left| \frac{m\, (e_m) + m\, (c_m)}{\sqrt{\dfrac{(N - n)^*\, var(e_m + c_m)}{N^*n}}} \right| <= t_{\alpha,n-1} \qquad (m = 1, ..., M)$$

$$t_l = \left| \frac{(N - m)^*\, var(b + d_l)}{\sqrt{\dfrac{m\, (b_l) + m\, (d_l)}{N^*n}}} \right| <= t_{\alpha,n-1} \qquad (l = 1, ..., L)$$

where *m (c)* and *m (d)* are the estimates of the means of the correcting components. The sample from which we shall have to evaluate the result is, however, of the same size and to be sure that the user’s quality requirements are met *the second set of criteria* must obviously be imposed:

$$\begin{array}{lll} var\, (e_m + c_m\,) <= & m\, (e_m\, ^*e_m\,), & (m = 1,..., \; M) \\ var\, (b_l + d_l) \; <= & m\, (b_l\, ^*b_l), & (1 = 1,..., \; L) \end{array}$$

These are the minimum restrictions any editing method has to satisfy if it should be applied. They can of course be applied independently to subgroups of the population.

144. Again we can observe a resemblance to the concept of required "accuracy” in the theory of sampling.[50] In the same way as the quality standard is necessary for the determination of optimal sample size, it is also necessary in order to decide on the optimal editing procedure. The impression of the author is that this very important aspect of the design of automatic editing procedure is rarely considered and that the numerical specifications are set independently of the quality standard

[49] C.D. Ferris, F.E. Grubbs and C.L. Weaver: *Operating Characteristics for common Statistical Tests of Significance*, Annals of Math. Statistics Vo1.17, 1946.

[50] W.E. Deming: *Some Theory of Sampling*, John Wiley and Sons, N.Y., 1950.

desired. In the US Bureau of the Census, the frequency o f corrections is used as a guard against inaccuracy in the results and as a measure of quality. The British Ministry of Agrlcu1ture, Fisheries and Food has also found that one of the most difficult problems encountered in automatic editing is that of devising an optimum level of acceptance.

145. Statistical observations usually comprise several variables, the required quality standards of which may vary. Therefore it is more relevant to think of the quality standard requirement as a set of probability statements of the above type. Even more complicated situations are met when quality standard requirements are related to time series Instead of a single measurement.

146. An examination of the descriptions of the different editing procedures used in the statistical offices Indicates that the majority relies on numerical specifications set subjectively by specialists. One important reason for subjectively set specifications is probably that the aim is, as is stated by the Department of Agriculture and Fisheries for Scotland[51], to maintain the general standard of accuracy previously attained by manual methods.

147. In foreign trade statistics where automatic ratio control seems to be usual, most applications reported so far are based on limits for the acceptance zone of the ratio specified by experts in the different trading fields. These 1imits are based on long experience, which has taught these specialists that ratios outside the limits are usually wrong and they are adjusted in the light of current information about general developments in the field concerned.

148. For correcting methods, the subjective specification is probably best illustrated by reference to the cold deck correcting method - even though the reasoning will be about the same for other correcting methods. The specification needed here is the statistical classification scheme and the representative records to be used. The details of the classification scheme are limited by the specialists ~ ability to distinguish between many types of statistical groups, In so far as the specialist is able to see any significant difference between two statistical sub-groups, he will try to specify them in his classification scheme. When he has obtained a classification scheme, which in his opinion expresses all significant differences in respect of automatic correction by means of a cold deck of representative records, his next task will be to specify the values of the representative records. An expert in the trade concerned will here be able to take into account much valuable outside information which is very difficult to introduce into the editing procedure by other methods - for example, recent developments in international trade relations, catastrophes, etc.

149. The subjective specifications are usually based on experience from previous surveys of the same type, but of course they may also be obtained from a sample of the statistical population to be processed. This makes it possible to adjust the experience in the light of fresh information from the population itself. One important principle, which is mentioned in the report from the British Board of Trade, is to fix the upper limits at no more than ten times the corresponding lower limits in order to pick up errors resulting from misplacement of the decimal point.

150. The main advantage of the subjective judgment specification compared with specification by statistical methods is that the former readily allows information of a non-statistical nature to be taken into account.

151. On the other hand, it is impossible to relate a quality standard requirement to a subjective specification and it is impossible to give any a priori objective statement about the degree of

[51] United Kingdom: *Report on Electronic Data Processing*, 1962 (Conf.far.Stats/WG.9/3 5/Add. 14)

accuracy achieved by using an editing procedure based on subjective specification. In other words it is difficult to evaluate the procedure objectively. In addition, the time needed for subjective specification and the 1ack of specialists may be a serious inconvenience as the following quotation indicates: [52]

"*It takes a considerable time until the necessary indications on admissible and non-admissible combinations and code numbers are available. In some cases the material could be produced only by the use of electronic computers*."

4.4.3 Statistical specification methods

152. The term statistical specification methods will be used to denote all specification methods used in editing which are based on the theory of mathematical statistics. The statistical specification methods are quantitative and thus permit the application of computers already in this preparatory stage of editing, and also the utilization of more extensive information. The computers will, for example, make it possible to distinguish among more groups in the classification scheme of the cold deck method.

153. The statistical methods require at least a random decision sample of records, which can be assumed to be drawn from the same population as those to be edited, and on which the necessary calculations can be performed. The specifying calculations may either be performed directly on the observed values of the sample or on thoroughly edited records giving both observed and true values.

154. The first approach is mentioned in several documents. for example in the French report,[53] the British report,[54] and in the recent explanatory memorandum on the lists of points to be covered in national papers on experience of using electronic data processing for statistical purposes (ECE document ME/31/62). The measures calculated from observed values are frequencies and measures of dispersion such as the standard deviation, range and percentiles. For a ratio control, it is, for example, necessary to specify the limits numerically. Depending on how many records we are willing to reject for inspection or automatic correction, the limits can be determined according to relevant percentiles of the sample or assuming an approximate normal distribution, She standard deviation of the sample multiplied by +k is used as a limit around the sample mean.

155. In correcting by the functional method, the correcting functions can be estimated from a sample of records and the regression function used in the subsequent editing.

156. The limits and the correcting functions specified in this way will be *stochastic variables* as they are based on a sample which is supposed to be random. The quality of the editing procedure will therefore depend on the sampling distributions of these variables or, implicitly, on the sample size, which has to be determined subject to quality considerations.

157. While the first approach applies statistical specification methods based on the assumption that it is satisfactory to gain Information about the distribution of the observational variable, the second approach requires Information about the joint distribution of the observational and the true

[52] Federal Republic of Germany: *Report on Electronic Data Processing*, 1962 (Conf.Eur.Stats/WG.9/35/Add.2).

[53] Institut National de la Statistique et des Etudes Economiques: *Report on Electronic Data Processing*, 1962 (Conf.Eur.Stats/WG.9/35/Add.9).

[54] United Kingdom: *Report on Electronic Data Processing*, 1962 (Conf.Eur.Stats/WG.9/35/Add. 14).

variables. This means that the sample first has to be edited by means of an expensive and nearly ideal editing procedure which makes 1t necessary also to consider the cost component explicitly.

158. According to the second approach, the limits of the ratio control ought to be determined on the basis of both the estimated true and the observed distributions, because efficient control must be concentrated on those fractions of the distribution space which are associated with the errors which have the largest effects on the results and this can only be learned by inspecting both the observed and true values. The correcting on the other hand ought to be based on the true distributions in order to obtain the true structure as a basis for producing corrected values.

159. The general editing model for this situation is Characterized by the conditions of the solution of the following problem in which the arguments $s_l, \ldots, s_{ls}$ specify the parameters o f the control and correcting methods:

$$Q\ (s_l, \ldots, s_{ls}) = min$$

subject to the first set of criteria:

$$t_m\ (s_1, \ldots, s_{ls}) <= t_\alpha$$

and to the second set of criteria:

$$v_m\ (s_1, \ldots, s_{ls}) <= v_m{}^0 \qquad (m = 1, \ldots, M)$$

Even with a restricted class of editing methods, the solution of the above minimum problem is formidable. It might, however, be approximately solved by the following procedure.

160. Before discussing the specification of the control zone, we shall look into some problems of specifying correcting methods since the specification of the control zones depends on the correcting method selected.

161. The restriction on the *var (m (e) + m (c))* implies a condition which any correcting method must satisfy:

$$m\ (c*c) + 2m\ (c*e) <= 0$$

or expressed by the correlation coefficient between c_i and e_i:

$$R\ (c*e) <= \quad \frac{1}{2}\sqrt{\frac{m\ (c*c)}{m\ (e*e)}}$$

The condition imp1ies that any correcting method must be designed in such a manner that the correlation never becomes positive and larger than the right-hand side of the expression.

162. The two restrictions on *t*-variable and variance together yield the second condition of the correcting methods which says that:

$$| m (e) + m (c) | <= | m (e)|$$

or squared

$$m (c)^2 <=2m (e) * m (c)$$

163. To exemplify the above stated principle we may select the very simple correcting method assuming that all units classified in the acceptance zone are correct:

$$c_i = -m (e)$$

All values of units in the automatic correcting zone are thus subtracted by the value of the estimated mean error with opposite sign. The variance and the covariance will both be zero and therefore the first condition is met independently of the size of $m (e*e)$. As to the second condition we can see that $m (c) = - m (e)$ and also the second condition is only just satisfied, A similar investigation of applied correcting methods will probably reveal that not all will meet the above conditions.

164. Linear correcting methods seem to represent an easily applicable class of correlating methods to which reference is frequently made. Let us consider the following correcting model:

$$c_{im} = A_0 + \Sigma_{k=1}{}^M A_{mk}*x_{imk}' \qquad (m = 1,..., M)$$

which implies that the correcting variable may depend on the size of one or more observed variables of the unit the m^{th} variable of which requires correction. Let us consider the particular numerical specification of the *A*-parameters which give as a result $m (c_m) = - m (e_m)$ and the least $var (e_{im} + c_{im})$ as the best, This is obviously equivalent to using the regression coefficients of $-e_{im}$ to x_{ij}', $(j = 1,..., M)$.

165. We turn to the second problem of specifying the limits of the three proposed zones, the manual inspection zone, the automatic correcting zone and the acceptance zone. This can only be solved by an exact method if it is possible to express the frequency function by the classifying characteristics as an analytic function from a limited class, which is rarely the case. We approach the problem by defining G *editing* groups such that each statistical unit is classified in one and only one group by means of some convenient characteristics. The control specification problem is now regarded as the problem of assigning each of these groups to one of the three zones. When the assignment has been completed, the limits of the zones are also specified by the definitions of the editing groups which each of them comprise. This implies that all units belonging to the same group must be equally treated, i.e. either manually inspected, automatically corrected or accepted.

166. By the conditions for the correcting methods we are sure that the variance restriction is always satisfied. The *t*-variable restriction has therefore to be taken care of in the assignment process. We first square and transform the restriction in this manner:

$$m (e)^2 + 2m (e)*m (c) + m (c)^2 - t^2\alpha* var (m (e) + m (c)) <= 0$$

and by the approximation

$$(m (e) + m (c))*(m (e) - m (c)) = m (e)^2 - m (c)^2 \sim (m (e) + m (c))*(a - 2m (c))$$

(where *a* is the error limit specified by the user in his quality requirement statement) we obtain:

$$-a^*m\ (c) + t^2 \alpha^* var\ (m\ (e) + m\ (c)) <= a^*m\ (e)$$

for each variable, $(m = 1, \dots, M)$.

167. The above restriction can be rewritten In a linear form in the binary integers E_{1g}, E_{2g}, and E_{3g} which represent the three alternative actions of manual inspection, automatic correcting or acceptance respectively. Recalling that in each group only one of the three members can be non-zero, the following formulation is valid subject to the approximation of $m\ (e) = a+m\ (c)$:

$$\Sigma_g^G (a_{1gm}{}^*E_{1gm} + a_{2gm}{}^*E_{2gm} + a_{3gm}{}^*E_{3gm}) >= a_{0m} \qquad (m = 1, \dots, M)$$

$$E_{1gm} + E_{2gm} + E_{3gm} = 1 \qquad (i = 1, \dots, 3)$$

$$E_{igm} = (0 \text{ or } .1) \qquad (g = 1, \dots, G)$$

$$a_{igm} = N_g [-a^*m_g(e_m) + t^2\ N_g\ var\ (m_g(e_m) + n_g(c_m))/N\ ]/N$$

where variance will be zero for a_{1gm} and equal $m_g\ (e^*e)$ for a_{3gm}. When the correcting method is determined, these are known constants.

168. By assuming a linear *Q*-function we are now able to express zone specification as the solution of the integer linear programming problem with 3^*G variables and $1+G$ restrictions, which may in principle, be solved by linear programming techniques:

$$Q = \Sigma_i^3 \Sigma_g^G\ g_{ig}{}^*N_g{}^*E_{ig} = max$$

subject to:

$$\Sigma_i^3 \Sigma_g^G\ a_{ig}{}^*E_{ig} <= 0$$

$$\Sigma_i^3\ E_{ig} = 1$$

where E_{ig} is *0* or *1*.

169. For each of the *M* variables an assignment solution must be computed according to the above set-up. It may be realistic to assume that lf a unit is classified in the manual inspection zone for one variable, the whole unit should be manually inspected. This can be done by defining an overall manual-inspecting zone. The whole assignment problem can be solved simultaneously for all variables by introducing all *U* quality restrictions in the programming model and setting all $E_{1gm} = E_{1g}$ which now gives a model in $2GM + G$ variables and $G + M$ restrictions.

170. This approximate solution will also give the basis for calculating the estimates t_m and for proving that the solution really meets all requirements, the *t*-estimates are also indicators of the accuracy of editing. The smaller the t_m -value, the better is the editing for this variable.

4.4.4 The decision sample

171. The need for empirical knowledge in general and statistical knowledge, in particular, about the kind and nature of errors, has been mentioned and a decision sample has been proposed as a mean for obtaining this knowledge.

172. Whether the sample is drawn from a previously collected set of data which is assumed to be representative of the data to be processed, or the sample is drawn from the data to be processed, the modern theories and methods of sampling should be applied as a basis for collecting the necessary information.[55] The size of the sample should as indicated in the section concerning quality standard to be determined with respect to the quality requirements by means of the operating characteristics curves.

173. Usually a sample will not be collected only to satisfy the need for knowledge about errors. This aim may only be one among several which are the basis for a *pilot survey*. The observations of this sample should be edited as thoroughly as possible for study and analysis of the kind and nature of the errors. By utilizing a pilot survey also for obtaining information about the errors. the most efficient editing criteria can be specified numerically 1n advance of the main Investigation and considered as an Integral part of the whole plan for the investigation.

174. When a pilot survey cannot be performed and data from a similar previous statistical investigation is available, a decision sample from this may be used. A particular possibility within this approach is the *post-enumeration surveys* often performed in connection with large surveys and censuses in order to evaluate the quality of the statistical results. By minor extensions of the scope of the post-enumeration survey, the needs of the editing for knowledge of the kind and structure of the errors may be satisfied. This knowledge gained by a post-enumeration survey analysis can often prove to be of great value in designing and specifying the editing of a future investigation.

5. EMPIRICAL RESEARCH

175. In a recent paper, Tore Dalenius writes:[56]

> "*For a long period to come, we may have to live with imputation techniques having rather weak theoretical foundations. It should therefore prove valuable to analyze the performances of different techniques under realistic (or nearly realistic) conditions. Two examples of feasible approaches will be given.*
>
> *(i) ………..*
> *(ii) Obviously, analogous experiments may be carried out by means of artificial sampling ("Monte Carlo").”*

The second and quoted approach has already been tried In the Norwegian Central Bureau of Statistics and one of the reasons for such an empirical approach is to verify methods developed on a theoretical basis.

[55] M.H. Hansen, et a1.: *Sample Survey Methods and Theory*, Volumes I and II, John Wiley and Sons, N.Y. 1953.

[56] Tore Dalenius: *Automatic estimation of missing values in censuses and sample surveys*, Statistical Review, No.7, Stockholm, pp. 395-400.

176. Another reason for intensifying performance research is the large investment in preparatory work connected with automatic editing procedures. This extensive and time-consuming investment makes it desirable to find editing methods which are applicable, not only to a limited, special class of jobs, but to large, general classes of jobs and which could be included as standard bricks in statistical processing. This need is pointed out by Thor Aastorp and the author.[57] A similar point of view is expressed in a recent document prepared by the Statistical Office of the United Nations, which proposes that a Master Programme might be designed to meet ordinary statistical requirements. One sub-routine which such a Master Programme must be able to perform is "verifying and editing data in standard form.[58]

177. Another reason for empirical research is that several editing problems may generate rather difficult mathematical problems if they are handled analytically and these problems may be avoided lf numerical methods are applied. Compared with a pure theoretical approach, this empirical research is therefore considered by the author to be more promising.

5.2 Research by simulation[59]

178. The basic idea of the scheme presented here is that instead of setting up a model of mathematical equations which have to be solved for various distributions, models of alternative actual situations are set up and the behavior of these models is studied numerically.

179. Different editing methods have for a long time been studied in order to evaluate which might be suitable for inclusion in standard programs. The assumption for an evaluation of this type is that we know the effect of the methods on different populations and are able to compare the edited values with the true values.

180. It will obviously be time-consuming and very expensive to obtain data comprising both observed and true values for the necessary populations which have to be included in the performance evaluation. The scheme proposed here is simulating the real production of true values of different population types as well as the mechanism generating errors by means of a computer. The specifications of the simulated mechanism can be changed experimentally to produce pseudo-records containing observed and true values covering different types of populations which are found worth while studying.

181. The editing procedures which are being studied are applied to the generated representative populations, simulating by computer also any manual operations if necessary. The results are then analyzed and evaluated by means of the computer.

182. The whole research scheme can be considered as composed of three parts which way be integrated in one computer application. The scheme is illustrated in *Figure 5*. Each time the computer starts on the first block a *run* is said to be started, and each run is equivalent to the collection of a set of records representing a new population and/or a new error generating

[57] Svein Nordbotten and Thor Aastorp: *On library routines in the statistical data production*, Statistical Review, No.31. Stockholm 1962.

[58] "*Master Programme for Statistical Compilation by Computer*", UN Economic and Social Council, Statistical Commission (Twelfth Session) Document E/CN.3/3O2, March 1962.

[59] This section is based on a paper prepared by the author for a meeting of the Scandinavian Committee on Technical Co-operation in April 1962.

mechanism. The number of runs is denoted by R in the diagram and all runs constitute a simulation *experiment*. Within each can we may have several applications of editing *methods* which we want to study, i.e. we may automatically apply several methods to the same data in order to be able to perform comparisons at a later stage. The number of methods applied in a particular run, r, is denoted by M which indicates that the number of methods which we wish to try may vary from one run to another within the same experiment. In the same way, the results of each applied method can be analyzed and evaluated according to different criteria. For the same method applied on certain populations, we may for example only want to analyze and evaluate the performance of the method as to the accuracy of certain statistics such as population totals or averages, while applied in other runs on other populations the performance as to the accuracy of the distribution characteristics may be important. The number of analyses within method application m_r, is denoted by A_{mr}

183. Each of the three parts of the simulation requires a computer routine comprising sub-routines representing the different generating mechanisms, editing and analysis methods. The relevant routines and specifications such as number of records, control limits etc. are compiled from parameters fed into the computer. Each of the three parts will be treated in more detail below and a preliminary simulation program worked out in Norway will be presented in the subsequent section.

164. A similar system has been described recently by R. Grimond. His system, which concerns data transmission over telephone lines, is also based on Monte Carlo generation of artificial error statistics and simulation of error-detecting schemes as a powerful way of assessing the effectiveness of the different schemes.[60]

165. The immediate objection to the idea of generating pseudo-observations is that they do not reflect reality. This is of course true, but the generation of pseudo-observations for this purpose is only one aspect and the least important. The derivation of the necessary statistics from actual records relating to a real population, or even applying the original records instead of generated observations would not cost much more than the generation,

186. The much more important and difficult aspect is that statistical offices do not have original records which contain *both* observed and true values which are essential in research of this type, A few offices will perhaps possess records with observed and edited values for some types of populations. But since the problem originally arose from the fact that an editing procedure never can give true values for all records, such data will of course favor editing methods related to the one originally applied and cannot be used, In addition one of our main aims is to find methods which have general good features. This requires also studies of situations not yet recorded, but which may occur in the future.

187. The generation requires for each run a specification regarding which out of alternative generating models should be used, e.g. whether the error components are additive or multiplicative, whether they are independent of or dependent on the sizes of the true values, whether the variables are linearly or non-linearly related to each other, etc. In the case of dependence between variables, this may either be exact or stochastic due to disturbance factors. This specification selects the correct generating sub-routine.

188. When the generating routine is se1ected, it will1 need specification about the dimension of the observations, i.e. the number of variab1es observed, the number of observations wanted, the

[60] R. Grimond: *An Analysis of Real and Simulated Statistics for System Design Purposes*, The Computer Journal, Vol. 5, No. 2, July 1962, pp. 94-99.

distributions of the true variab1es, error components and disturbance factors, parameters for re1ations between varlab1es, etc.

189. The generation is performed by the generation of a set of random numbers which are used together with the described distributions and relations to produce the n-dimensiona1 observed and the related true values. A simulated record thus is composed by these *2*n* values and is stored in a magnetic tape file called the *observation file*. There will be one observation file for each run of the experiment.

190. The second part of a simulation experiment of editing will be the method to be evaluated. The frame of the experiment requires that any method which is included or simulated must be modified to accept the data from the observation fi1e without of course utilizing the true values during the processing. All methods must also as a standard result give records with *3 *n* values which are the observed values, the true values and the edited values. These records are stored in a magnetic tape fi1e as the *edited file*.

191. The procedures for complete automatic editing do not need much modification to be included in the simulation scheme. Semi-automatic procedures requiring for example manual inspection of some rejected cards, must, however, be equipped with routines simulating these manual processes. This may often be a problem since these processes are left manual just because they are difficult to simulate by an automatic process. Objections may therefore be presented on the basis that the simulation applications will favor those methods, which are well suited for automation. To meet this objection the semi-automatic methods may deliberately be favored by pretending that manual operations always are correct. This is done by substituting true values for observed values while simulating a manual part of the editing.

192. The simulation experiment will also be a very powerful instrument for studying and evaluating the effects of different specification methods. In one run we may for example include a series of applications each of which involves a systematic change in the specifications of the acceptance zone. This is a very important application since most of the distributions we have to deal with are far from normal, and therefore difficult to handle in an analytical way.

5.2.3 The evaluation of the results

193. Each application results in an editing file and the main aim of the whole simulation experiment is to analyze and evaluate these files. The evaluation must of course be considered in relation to the quality standard requirement. This means that in some cases it will be relevant to consider the deviation of the estimated average from the true average as a measure of quality according to which the methods ought to be evaluated, while in other cases the main interest is attached to deviation measures between the distributions of edited and true values. One such measure is the χ^2-statistic calculated on the basis of the frequencies of edited and true variables in the classes of a pre-determined classification system.

194. From the point of view of general applicability, it may be desirable to study the resu1ts of a certain method, analyzed and evaluated by different measures. This analysis may for example Indicate that a particular editing method, say method *A*, may be very well suited for editing a very large class of populations if the quality standard requirement refers only to single statistics such as the total value or average. If, however, the quality requirement refers to statistics in sub-groups, i.e.

to some rough distribution characteristics, method *B* may seem to be in general the most adequate, or if it is expected that the data are going to be classified in a number of alternative ways, the requirement can indicate a deviation measure based on a large number of groups with the result that in this case method *C* is the best one in general. The analysis part of the simulation must therefore be controlled by selective parameters selecting the appropriate analysis.

195. The analysis and evaluation must also take into account the cost aspects of the different methods and should therefore give cost indicators as well as quality indicators. The cost indicator is a function of the numbers of records treated in different ways, for example the number of records selected for pilot sample, the number of records accepted, the number of records automatically corrected and the number manually corrected. The numerical specification of the cost indicator function depends on the relative costs of these different treatments and can be determined parametrically.

196. The result of each analysis is thus a set of evaluation coefficients representing cost as well as quality.

5.3 Application of a simulation scheme

197. The above simulation scheme has been programmed and applied in a simplified form in Norway and a description of the Norwegian program is given here as an illustration of the ideas set out above. The equipment available In Norway at the time of programming did not permit an integrated experiment program. Three independent programs are therefore used representing each of the three parts of the experiment.

198. The observation-generating program is only able to produce one- or two-dimensional observations. Each record therefore is composed of four values, x_1, x_2, x_1', x_2', of which x_2 and x_2' both are zero if only one-dimensional observations are wanted. The true variable x_1 has a distribution, which must be specified according to the situation, which we want to study. The specification is done by a table of *100* cells describing the cumulative distribution of x_l. This table is read in parametrically be fore generation starts.

199. Between the true values x_l and x_2 the following relation call be introduced:

$$x_2 = (a + b{*}x_1 + c.x_1^2)\ {*}d$$

where *a*, *b* and *c* are specified structural parameters while *d* in general is a stochastic disturbance with a distribution to be specified in the same way as the distribution of x_l. In particular this distribution may contain the same non-zero value in all cells, which is equivalent to the situation where *A* is an exact relation exists between x_l and x_2.

200. The observed values are generated by the following transformations:

$$x_1' = x_l\ {*}e_l$$
$$x_2' = x_2\ {*}e_2$$

where e_l and e_2 represent the stochastic errors and are values obtained by random drawings from a specified probability function. It should be noted that the errors are here assumed to be multiplicative.

201. For each record, four random drawings are necessary and just as many random numbers. Instead of generating random numbers this program reads from a large file.

202. To start generating, the program therefore needs the following parameters read from cards:

1) N: number of records to be generated.
2) a, b and c: structural parameters.
3) A (3 x 100) matrix on cards the rows of which describe the x-, d- and e-distributions. respectively.
4) A file of random numbers in a compressed binary form on cards.

These specifications permit the generation of a large class of different two-dimensional populations. In particular, the matrix representing the three distributions makes it possible to describe any distribution in practical detail.

203. Preceded only by the parameter *N*, the observation file is then punched out in a compressed binary form on cards, which are the standard to any part *2* programs.

204. The part *2* program varies in detail of course depending on the editing procedure it represents. Some standard rules must, however, be satisfied, i. e. the processing is always performed on the standard output of part *l*, and the result of any procedure is the edited file in a standard compressed binary form preceded only by *N*. Simulation programs for several of the editing methods discussed in this paper have been prepared.

205. The third program of the simplified simulation scheme is the analysis program. Several different forms of evaluation can be performed and the one wanted is selected by a selective parameter card. After this card and the necessary specifications have been read into the computer, the standard output of the editing programs is processed.

206. The evaluation forms comprise the two ratio coefficients:

$$t_{x1}'' = \Sigma x_1'' / \Sigma x_l$$

$$t_{x2}'' = \Sigma x_2'' / \Sigma x_2$$

and the correlation coefficients:

$$r_{x1''*x1xl} = m_{x1''*x1} / \sqrt{}(m_{x1''*x1''} *m_{x1*x1})$$

$$r_{x2''*x2} = m_{x2''} / \sqrt{}(m_{x2''*x2''}*m_{x2*x2})$$

The ratio coefficients are useful when studying the effect on totals and average while the correlation coefficients refer more to the effect on the classification of the individual units by edited values.

207. All evaluation forms include a very simple form of cost indicator calculations resulting in the coefficients p_l and p_2 which are simply the relative number of corrected values of variable number one and two. These coefficients are those applied in the US Bureau of the Census as a guard against inaccuracy, while here against expensive methods, which may be regarded as two aspects of the same thing.

208. As is understood, the above set of programs is a very elementary example of simulation experiments, but work with such a simplified scheme will probably give experience about how to construct the more general, complex and integrated system necessary for systematic research in editing methods.

209. This system is being used to study the effect of some of the editing methods described in section 3 subject to varying specifications of distributions deduced from decision samples. To illustrate the approach a few results from a simplified experiment are described in the following paragraphs.

210. The experiment reported was based on the following generation specifications:

1) *N: 100 records to be generated. This is a too small number to enable significant conclusions to be drawn about the methods applied. The conclusions only refer to the mass of 100 records itself.*
2) *Distribution: As illustrated in Figure 6, the F_l (x) - distribution corresponds to the typical skew distribution observed in many statistical populations. The F_2 (d) - distribution is specified as normal, (1.00, 0.05), while F_3 (e) - distribution is a discreet error distribution, which implies that 25% are error-free, IQ% are unspecified values etc. The coefficient b=1, the other two are zero.*

211. The following four editing methods were among those studied by using data generated in accordance with the above specifications:

ml) Ratio control method. This control method was combined with a new and exact observation of all rejected units simulated by substituting observed values by true values. The control 1imits required by the editing method were assumed to be fixed subjectively on the basis of knowledge about the standard deviation *s (d) = 0.05* and the mean *m (d) = 1*. The limits are specified as *m (d) = 3*s (d)*:

$b_l = 1- 0.15 = 0.85$

$b_u = 1 + 0.15 = 1.15$

m2) The hot deck correcting method. This correcting method is combined with the above described control method, Both values of a rejected record were in accordance with the hot deck method, replaced by the observed values of the last accepted record. The control specifications correspond to those used in *m1*).

m3) The gross error control method. This control method was combined with new and exact observation of all rejected units simulated in the same way as in *ml*), The control was performed on the ratios between x_1^1 and x_2^1 from samples of three and three successive records, the control is such that the risk of rejecting a correct record world be *5%* if the ratios were normally distributed.

m4) The non-editing method. This is not an editing method at all as all observed values are accepted, but a conversion from the form in which data are recorded in the observation file to that in which they are recorded in the edited file. The reason for including this procedure in the experiment is of course to be able through the analysis and evaluation phase to evaluate the necessity for editing at all.

212. The results of each method application were analyzed and the ratio and correlation coefficients as well as the relative number of corrected values were computed. The values of these coefficients are given in the following table:

Table: Results of a simulation experiment.

Method:	t_{x1}	t_{x2}	r_{x1}	r_{x2}	P	P_2
m1	1.001	1.005	0.999	0.999	0.34	0.36
m2	0.961	0.997	0.589	0.576	0.41	0.41
m3	0.845	0.930	0.867	0.918	0.10	0.12
m4	0.796	0.854	0.826	0.885	0.00	0.00

The last row of the table indicates that the observed values give biased results which understate the totals or averages of the mass.

213. The figures referring to the gross error control method show that this method changed about *10%* of the observed values while the corresponding figures for the ratio control method were about *35%*. On the other hand the *m3*) did not eliminate more than about one third of the biases while *ml*) eliminated them completely and also gave high correlation coefficients.

214. As to correcting methods, a comparison between *ml*), *m2*) and *m4*) indicates that the hot deck correcting method eliminated the biases almost completely, but deduced the correlation coefficients and correct classification of the individual observations.

215. The results of this simplified experiment have of course no general validity and must only be treated as an illustration of the described research approach.

6. FUTURE WORK ON AUTOMATIC EDITING

216. Several statistical offices have emphasized in their reports on electronic data processing the importance of automatic editing and the need for further work in this field, particularly on automatic correcting. The present paper gives a survey of editing methods, which have been

applied or proposed. When taking into account the variety of populations and their extent, relatively few methods have been developed so far and one of the main tasks of the future will be to increase the arsenal of useful methods.

217. A reasonable development of editing methods requires1 however, that the nature and objectives o f the editing be clearly specified and understood, i.e. a theory of editing is needed. Much work is required if editing theory is to be as good a guide to the statistician as the quality control theory is to the industrial production controller.

218. The automation of statistical processing is probably in its beginning stages. To obtain full advantage of automatic equipment it will be necessary to integrate operations and find the implications of automatic editing on other statistical operations and vice versa.

219. In the US Bureau of the Census, computers are used for automatic coding of industry codes in censuses and surveys of manufacturers and similar applications are also reported from Netherlands Central Bureau of Statistics. This operation seems often to be integrated with editing in manual processing and may influence the solution of the automatic editing.

220. It is envisaged that because of the increasing need for complex statistical information, actual as well as historical information must be kept in readiness on computer media files. The question whether data should be edited before being filed for general purposes or each time that they are retrieved for a particular purpose is an interesting as well as a difficult one with implications for both editing methods and file organization.

221. The use of advanced description of the statistical population by means of moments of higher order, regressions, correlations, etc., is also increasing with the introduction of computers in statistical work. It is very likely that the use of these measures will create new requirements which editing methods will have to satisfy.

222. All the above-mentioned problems are omitted from the discussions of the present paper, but are important questions needing answers. It is therefore hoped that in the future methodological work on automatic editing will be found to deserve more attention from statisticians.

Position:	Value:	Code:
1	0	14
2	0	15
3	1	16
4	0	17
5	1	18
6	0	19
7	0	20
8	0	21
9	1	22
10	1	23
11	1	24
12	0	25
13	1	26
14	0	27
15	1	28
16	1	19
17	0	30
18	0	31
19	1	32
20	1	33
21	0	34
22	0	35
23	1	36

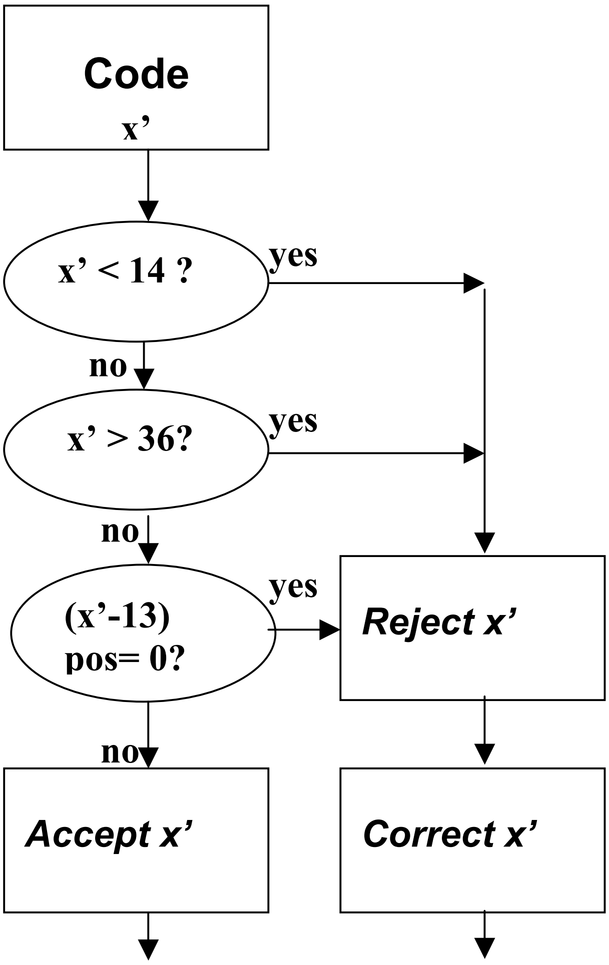


Figure 1: Code control by binary string technique

x_1' Value of production:	00-09	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
9000-9999										99
8000-8999									88	
7000-7999								76	77	
6000-6999							65	66	67	
5000-5999						54	55	56		
4000-4999				42	43	44	45			
3000-3999			31	32	33	34				
2000-2999			21	22	23	24				
1000-1999	10	11	12	13						
0000-0999	00									

x_2' Employees

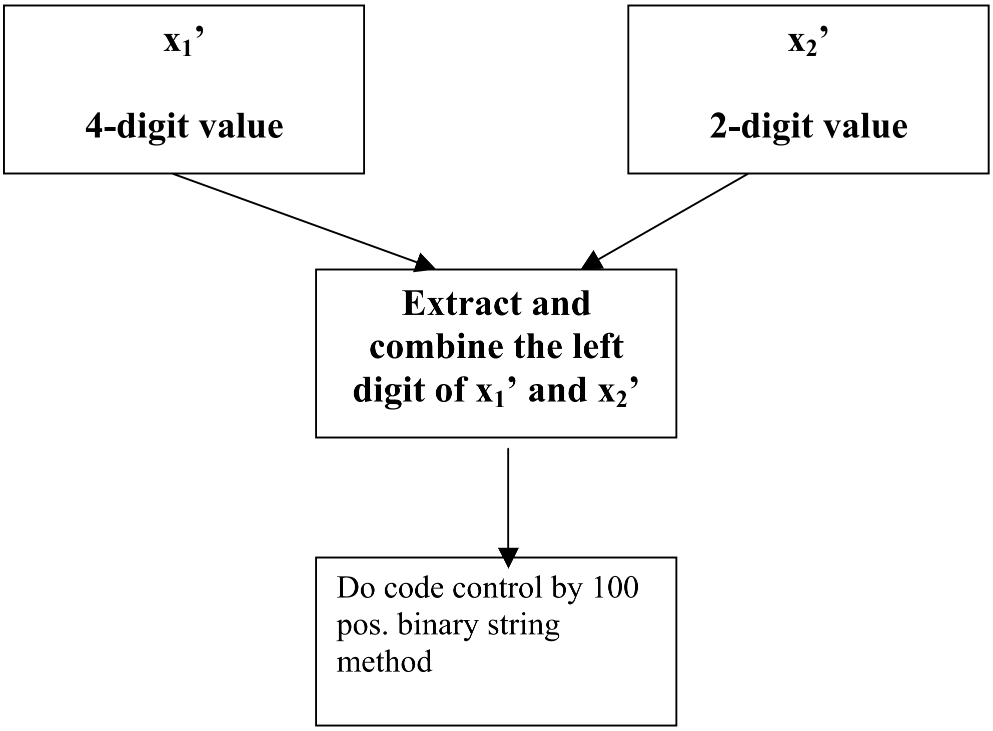
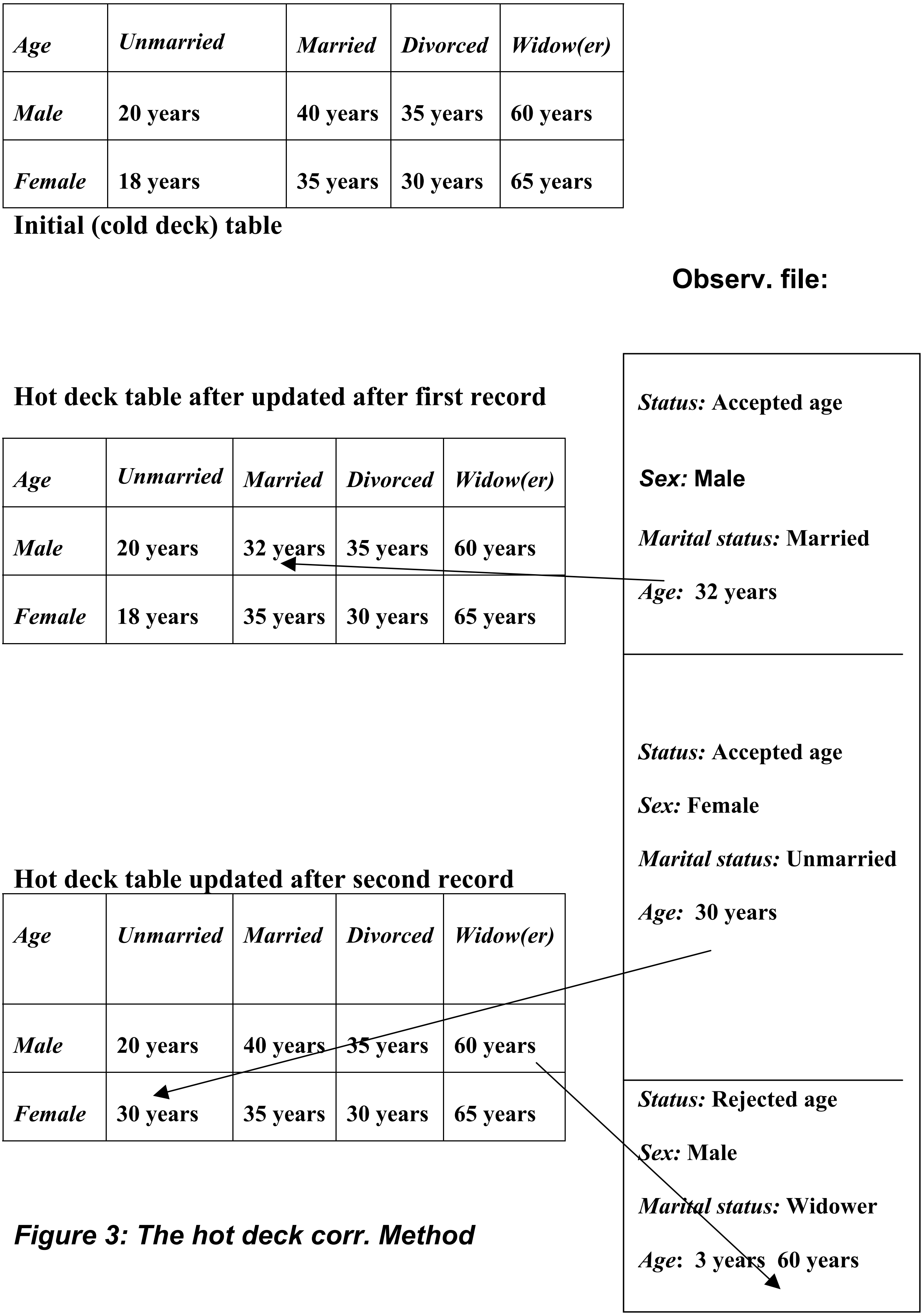

x_1'
4-digit value
x_2'
2-digit value
Extract and combine the left digit of x_1' and x_2'
Do code control by 100 pos. binary string method

Figure 2: Code control of economic variables



Age	*Unmarried*	*Married*	*Divorced*	*Widow(er)*
Male	20 years	40 years	35 years	60 years
Female	18 years	35 years	30 years	65 years

Initial (cold deck) table

Observ. file:

Hot deck table after updated after first record

Age	*Unmarried*	*Married*	*Divorced*	*Widow(er)*
Male	20 years	32 years	35 years	60 years
Female	18 years	35 years	30 years	65 years

Status: **Accepted age**

Sex: **Male**

Marital status: **Married**

Age: **32 years**

Status: **Accepted age**

Sex: **Female**

Marital status: **Unmarried**

Age: **30 years**

Hot deck table updated after second record

Age	*Unmarried*	*Married*	*Divorced*	*Widow(er)*
Male	20 years	40 years	35 years	60 years
Female	30 years	35 years	30 years	65 years

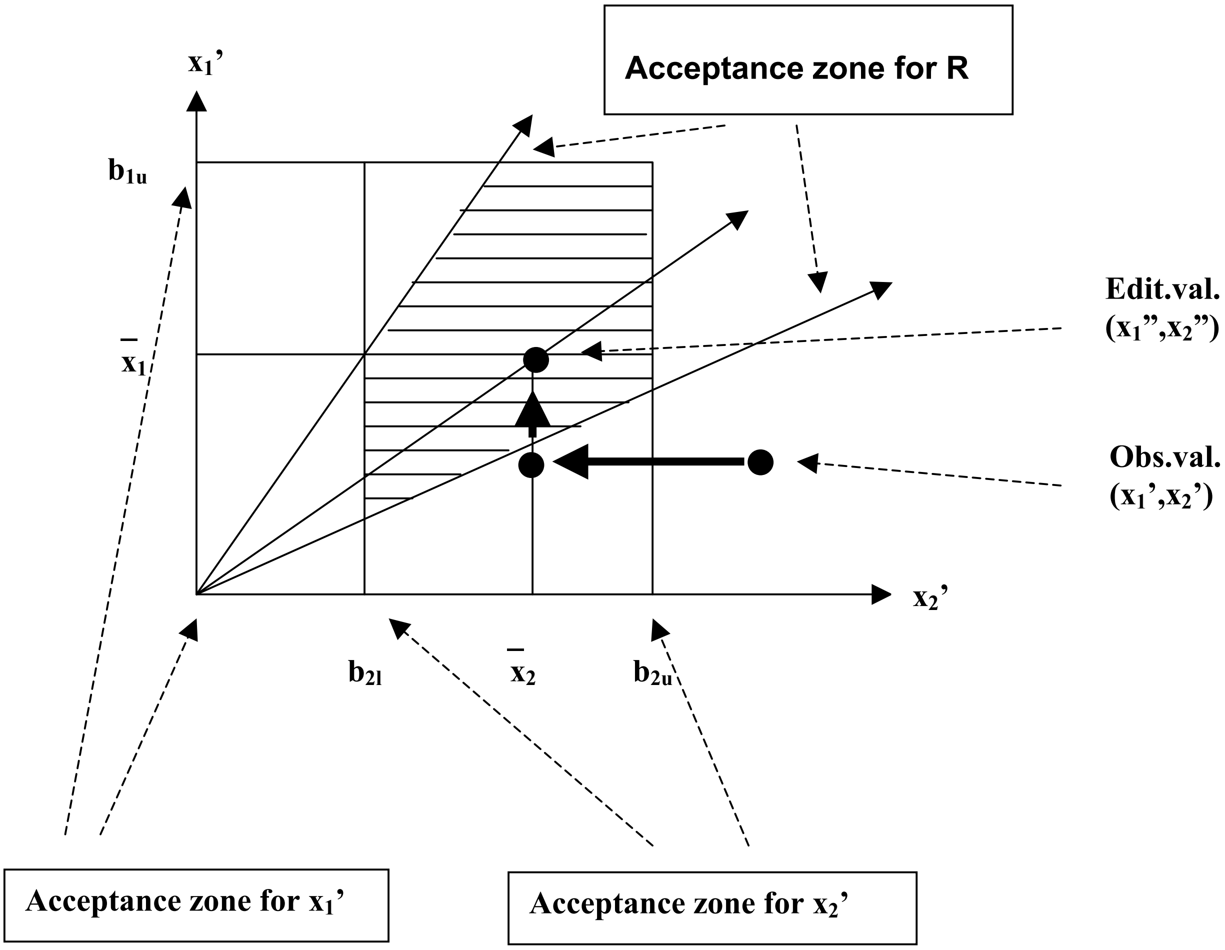
Status: **Rejected age**

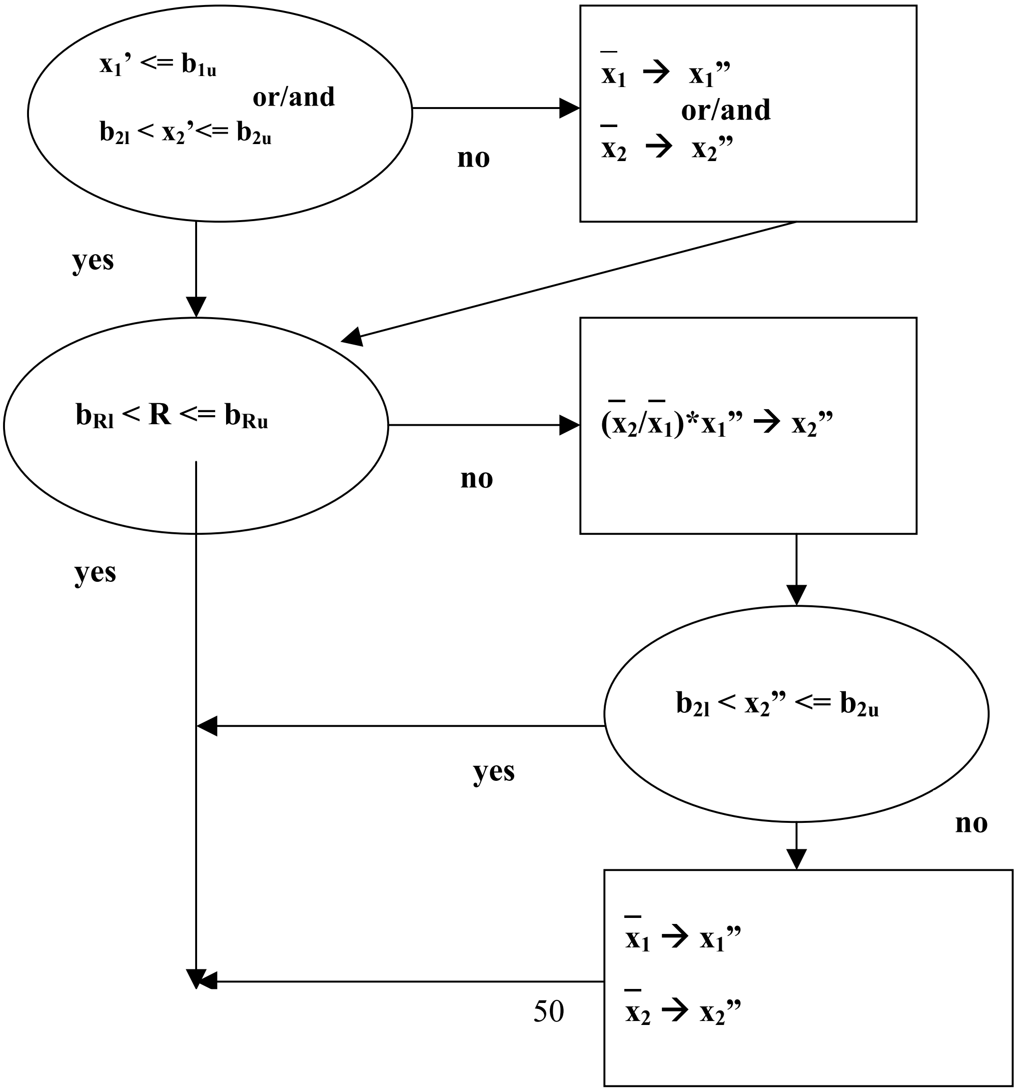
Sex: **Male**

Marital status: **Widower**

Age**: 3 years 60 years**

Figure 3: The hot deck corr. Method


Acceptance zone for R
x_1'
b_{1u}
$\bar{x}_1$
Edit.val.
(x_1'', x_2'')
Obs.val.
(x_1', x_2')
x_2'
b_{2l}
$\bar{x}_2$
b_{2u}
Acceptance zone for x_1'
Acceptance zone for x_2'


$x_1' <= b_{1u}$
or/and
$b_{2l} < x_2' <= b_{2u}$
no
$\bar{x}_1 \rightarrow x_1''$
or/and
$\bar{x}_2 \rightarrow x_2''$
yes
$b_{Rl} < R <= b_{Ru}$
no
$(\bar{x}_2/\bar{x}_1)*x_1'' \rightarrow x_2''$
yes
$b_{2l} < x_2'' <= b_{2u}$
yes
no
$\bar{x}_1 \rightarrow x_1''$
$\bar{x}_2 \rightarrow x_2''$

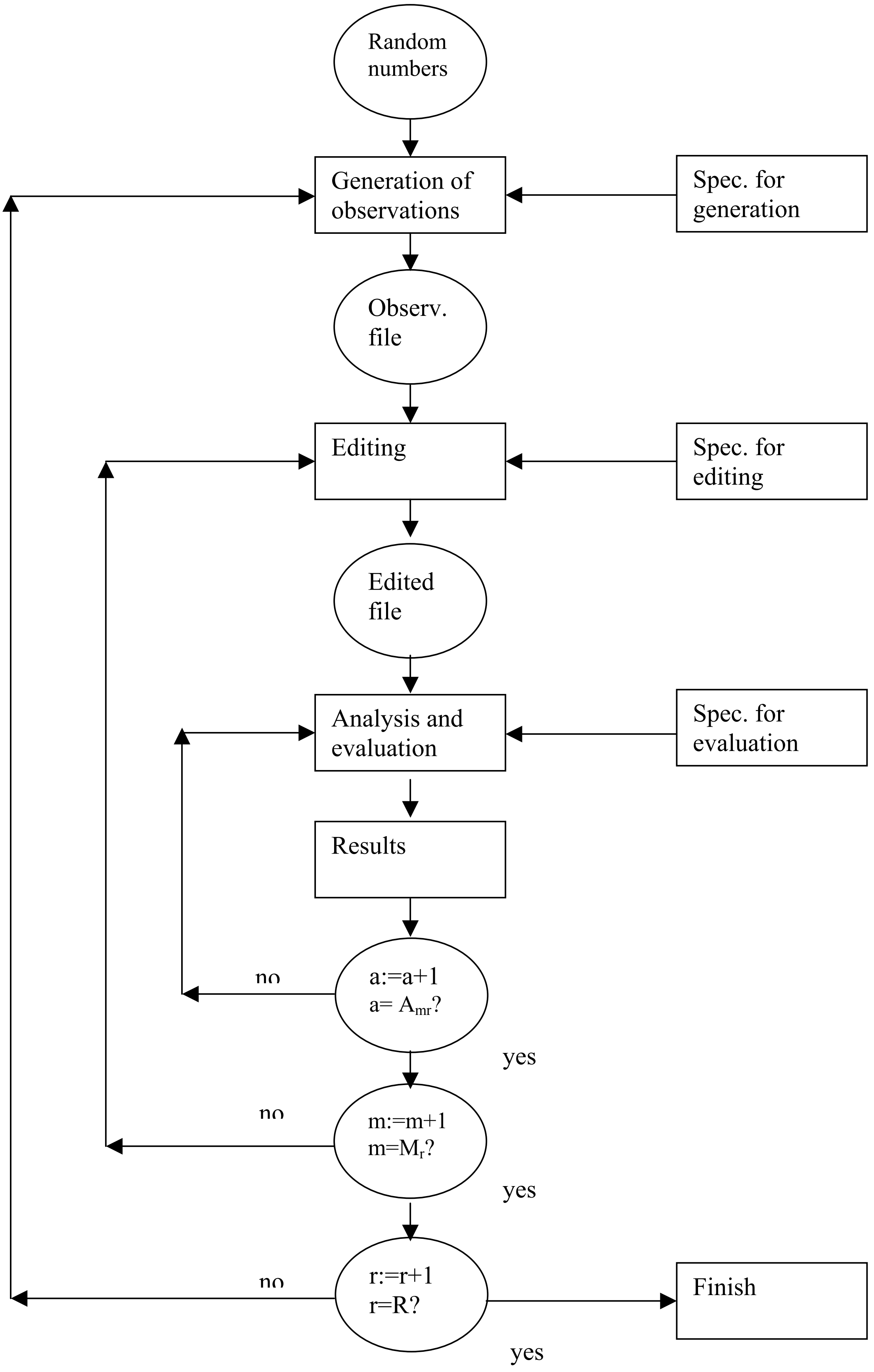


Figure 5: The successive stages of an experiment

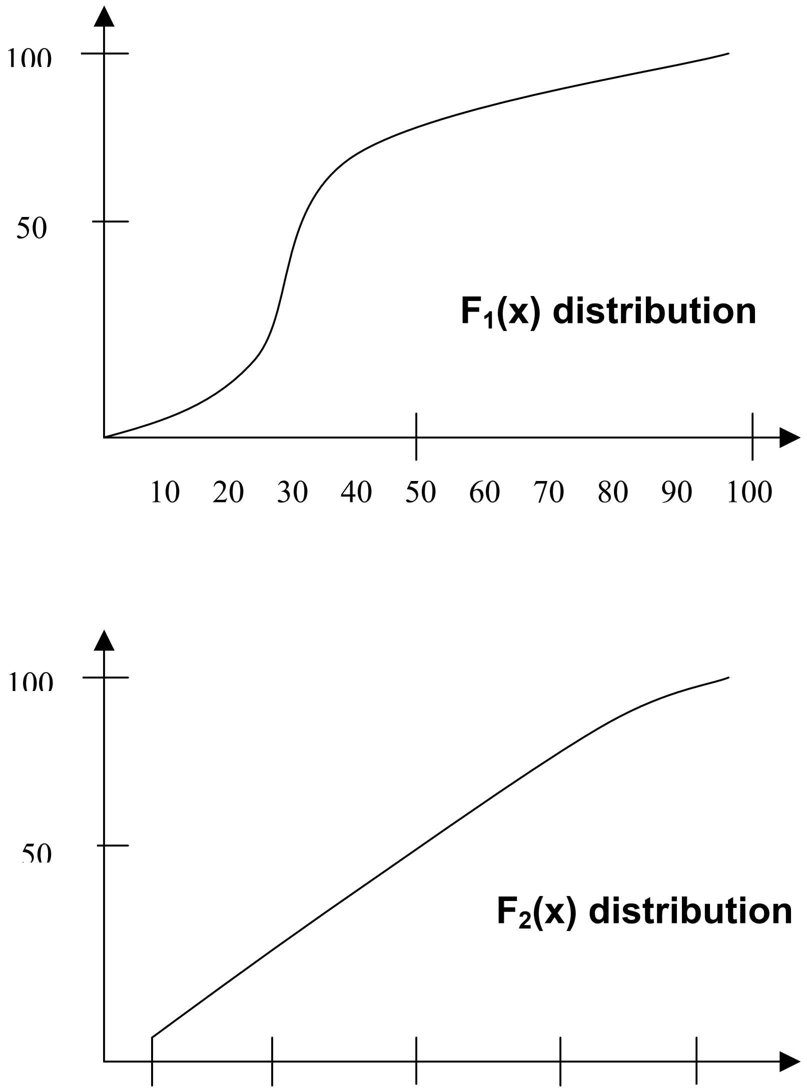


Figure 6: Distributions used in the experiment reported in section 5.3