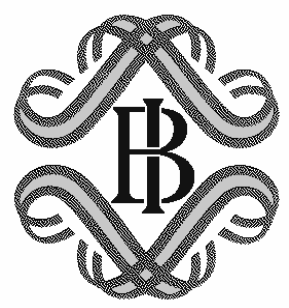
BANCA D'ITALIA

Statistical data and concepts representation

Vincenzo Del Vecchio



September 1997
(English version 2001)

The views expressed are those of the authors and do not involve the responsibility of the bank

INDEX

Summary

This work addresses the issue of the models for representing statistical data and concepts.

Chapter I accordingly goes to the basics of descriptive statistics, discussing the notion of "statistical data" and introducing that of "statistical concept". Three types of information to represent are identified. First is that of statistical concepts, i.e. the abstract definitions of aspects of the reality under study. Second is the "intension" of the data, i.e. the summary definition of their structure and significance. Third is the data proper, i.e. their "extension".

The third category is the one that actually contains information on the system under study. The first two provide "meta-information," or information about other information, and it is on them that this work focuses in Chapters II and III.

The three categories are appropriately related. Each is essential but, by itself, insufficient. Only with all three can we obtain a full, consistent representation of statistical data.

Introduction

That information is a strategic resource in the operation of any enterprise goes without saying. Just as universally agreed on is the need for a clear, documented, managed architecture for the data.

The Bank of Italy naturally pays great attention to informational activity, in particular the collection, processing and dissemination of statistical data.[1] Over the years, the Bank has handled a steadily increasing volume and variety of data, as its institutional responsibilities have evolved and expanded. The administration of statistical data has proven to be of fundamental importance,[2] and the relevant methodological tools have begun to be studied, starting with models for representing statistical data.

There is an abundant literature on the usefulness of models for representing and administering "analytical" data. There are a number of types of model, some of which have become de facto standards, thanks to their dissemination worldwide, such as the "relational" model.[3]

For statistical data, by contrast, representation models are much less common.[4] An especially rich school is the Swedish, which boasts the first theoretical theses on "infological" models in the 1960s and 1970s[5] and a long series of later elaborations[6] (some of which we shall refer to later on) designed to apply infological models to statistical data.

At the Bank of Italy, the last three decades have witnessed a series of practical applications and theoretical formulations in a process of positive feedback. The concepts derived from experience and the contributions of the literature (e.g. the conceptual statistical model[7]) have been applied and tested in action, providing new themes and ideas for research. The increasingly pressing needs of statistical data management and processing have prompted further theoretical refinements. The end result is an approach to modelling that is strictly correlated with the automatic processing.

Let us now summarize the principles followed in the Bank's automated data processing.[8]

Whereas the purpose of the extensional form of the data is to describe a part of reality, data definitions[9] have the purpose to describe the data themselves. The term "metadata" is used to indicate data describing other data. In statistical operations, memorizing data and metadata in the same information system can produce very substantial benefits.

First, the system's users can consult both the data and the metadata together, using the latter as guide to locate and interpret the data they are interested in.

---

[1] For an account of the Bank's statistical activities on credit and finance, see [2] [3], [11], [12], [15], [22], [34].

[2] By administration we mean, in particular, the planning, documentation and handling of statistical information flows to serve the needs of a large group of data users and suppliers, both inside and outside the Bank, and guaranteeing integrated, consistent development of the information system.

[3] See [13].

[4] Among those proposed are Johnson [23], Subject [39], Grass [36], SAM [Semantic Association Model) [41], CSM (Conceptual Statistical Model) [9], [17] or, in Italian, [6], [7].

[5] See [24], [25], [26], [27], [42], [43].

[6] See [30] , [31], [32], [44], [45], [46], [47].

[7] See [8].

[8] For a more thorough illustration of the Bank of Italy's approach, see [10], [14], [28], [37], [38].

[9] Data definitions are the result of applying a model to the modelling of a specific case (see [16], pp. 138, 140).

Second, appropriate "generalized" software,[10] which the Bank's Statistical Services Department[11] has been using for years now, can automatically (under the guidance of data and process descriptions), perform the processing necessary to constitute a complex statistical data base (filing, checking, calculation, consultation, statistical production, publications, etc.).[12]

These procedures guarantee the intrinsic consistency of metadata with data, thanks to the "active" role of metadata in data processing. The procedures are powerful and flexible in handling masses of statistical data with various structures and in constant evolution.[13] In most cases, the treatment of new data or the modification of previously processed data entail only the introduction or modification, within the information system, of the relevant definitions (with no need to alter software), and require very little time or expense.

Third, metadata can be defined directly by the data "administrators", whose activities become practically independent of applications developers. The metadata themselves can be automatically processed (to give help for their definition, to check their congruence, to disseminate them, and so on).[14]

This approach generated the "matrix" model, which while never completely formalized[15] has nevertheless been widely used both for the definition and administration of the statistical data and as guide for the elaboration of their "extensional forms".[16]

It gradually came to be realized that one particular category of elements, within the class of metadata, had peculiar characteristics: "statistical concepts". These are abstract definitions necessary to understanding the meaning of the data. A substantial body of such concepts has been built up over the years, generally called the "dictionary",[17] which has proven to be a fundamental tool for data administration.

The concepts, together with the definitions of the data and the actual data, are designed to construct a "full and internally consistent" representation, i.e. one that contains not only the datum but also the information needed to understand its structure and significance. In recent years, therefore, there has been special interest in developing a better formalization of the matrix model and its extension to the representation of statistical concepts.

Internationally, there has been increasing attention to modelling statistical data and, especially of late, an awareness of the importance of metadata. For years the UN Economic

[10] I.e. software capable of operating on data with different structures on the basis of the formal definitions of the structures and the rules for processing, memorized in appropriate files.

[11] This is the Department responsible for gathering statistics on the banking and credit system; it is referred to as SISC, the Italian initials standing for Servizio Informazioni Sistema Creditizio.

[12] The main procedure for gathering, filing and using data on credit and finance is PRISMA. Operational since 1989, this procedure now processes 34 periodical data sets, has some 700 million records in its files (now rising at the rate of over 100 million a year) concerning more than 40,000 different data definitions since 1984. The data base occupies over 100 Gbytes and serves more than 1,200 users in the Bank of Italy's head office and branches and in the Italian Foreign Exchange Office.

[13] Statistical data change more frequently than data connected with operational needs, mainly because users' information needs are more changeable.

[14] Within the Bank of Italy information system on credit and finance, there is a special procedure for metadata processing (the Metadata Administration System), created by the generalization approach just described.

[15] See [4].

[16] Apart from the PRISMA procedure, the matrix model, appropriately adjusted, has been utilized also in more recently created procedures, such as those for automatic production of publications, for handling the reports of the Central Credit Register, for bank account analysis models.

[17] It too is memorized together with the statistical data base and plays an active role in processing.

Commission for Europe has been engaged on this front.[18] In 1989, Eurostat began an initiative to define a standard for the interchange of statistical data within the framework of EDIFACT standards: the generic statistical message (GESMES).[19] The GESMES uses a conceptual model for the representation of statistical data permitting representation of the data and of their associated descriptions (metadata).

The "matrix" model in use at the Bank of Italy's Statistics Department is similar to and largely compatible with the GESMES conceptual model. The latter is designed for the transmission of "flows" of statistical data in electronic form and offers, for this purpose, a very broad spectrum of possibilities, whereas the matrix model is designed mainly for the administration and automatic processing of data memorized in the archives of a statistical information system.

The present paper traces a general description of the problem of representing statistical data and concepts as it has emerged from the formalization of the experiences of recent years in automating the Bank's information system on credit and finance. It presents the theoretical orientations and bases on which this activity is grounded, also with reference to comparable work in the international literature.

[18] See, e.g., [40], [49], [50].
[19] See [18], [19], [20], [21].

1. THE “STATISTICAL DATUM” AND ITS COMPONENTS

1.1 The "datum" in descriptive statistics

Descriptive statistics is a method of inquiry for producing a knowledge of reality. Its purpose, in fact, is to describe phenomena relating to a set of elements.[20]

The fundamental presupposition of statistics is that the inquiry is performed on sets of elements called "statistical units"[21] and that these elements can be described by means of attributes that may be called "characteristics".[22]

Statistical units may be any sort of objects or events, real or virtual, such as natural phenomena (precipitation), the executions of a measurement program, dice rolls, people (e.g. Italian citizens), groups of people (households, for instance), social and economic phenomena (marriages, employment relations, bank transactions), and so on.

"Characteristics" will differ in accordance with the statistical units under consideration. For individuals, meaningful characteristics could be gender, income, occupation; for "bank transactions" they might be type of transaction, name of bank, amount, currency, maturity.

Characteristics occurs in "modes" or "modalities".[23] That is, the characteristic "gender" has the modes "male" and "female"; the characteristic "income" occurs in a positive integer.

The whole purpose of statistical inquiry is to generate the "statistical datum," which expresses the modes in which a characteristic relating to groups of statistical units occurs. The datum thus expresses a type of property (the characteristic) relating to groups of objects.

The statistical datum is obtained by forming groups of statistical units according to criteria that depend on the purposes of the inquiry, observing the characteristic one is interested in and assigning to every group the mode that occurs. For example, suppose the statistical units are four people named Marc, Ann, Joe and Mary and that we are interested in knowing the "total income" for the following groups:

Group 1: {Ann, Joe}

Group 2: {Ann, Mary}

Group 3: {Marc, Joe, Mary}

the statistical datum is the rule that assigns to each of the three groups their respective incomes:

Group 1: 45 million lire;

Group 2: 34 million lire;

Group 3: 57 million lire.

[20] The concepts set forth here are available in elementary statistics textbooks. We have used terminology drawn from the introductory sections of [33] and [35].
[21] See [33], pp. 9-10.
[22] Ibid., p. 5; [35], p. 25 ff.
[23] See [33], p. 11 ff.

That is, the existence of three elements is postulated: a set of groups of statistical units, or "grouping"; a "characteristic" that refers to the elements of the grouping; and a "rule" that associates a mode of the characteristic to each set of the grouping.

The characteristic, obviously, must be meaningful for each set of the grouping and can be either quantitative (as in the case of "income") or qualitative[24] (for instance, a judgment on the wealth of the groups, expressed via the modes "rich", "well-to-do", "poor").

In general, the greater the extent to which the statistical units are grouped according to a criterion of homogeneousness, the more significant the resulting datum. Usually, homogeneousness is attained by expressing the criterion for grouping in terms of appropriate characteristics of the individual statistical units.

For example, considering the statistical units of "individuals", it may be useful to form groups of persons who are homogeneous in terms of gender and age. In this case, the combinations of modes of the characteristics gender and age (which are set a priori) identify the groups that we intend to observe. That is

Group	**Gender**	**Age**
group 1	Males	Young
group 2	Males	Old
group 3	Females	Young
group 4	Females	Old

The statistical datum appears as follows:

Gender	**Age**	**Income**
Males	Young	150 million
Males	Old	270 million
Females	Young	130 million
Females	Old	240 million

Thus the characteristics can be used in two ways: as the datum resulting from the inquiry or in order to identify the sets of the grouping. The former are called "statistical"; the latter, "systematic".[25] "Statistical" characteristics are, a priori, unknown; that is, there is uncertainty over which mode will occur; "systematic" characteristics are determined, a priori, by the observer (for instance, if one wants to study how the average income of Italians varies with age, average income is a statistical characteristic, as one does not know how it will occur beforehand, while age is systematic, as its values are predetermined).

The foregoing considerations abstract from the "temporal" dimension of phenomena. Yet in their own experience, people intuitively perceive the evolution of facts and events over time. The

[24] See [33], pp. 11-15; [35], pp. 28-29.
[25] See [33], pp. 227-230.

"time" factor is thus essential to statistical description. In particular, the properties of groups of statistical units (that is, the occurrences of the statistical characteristic) may vary over time.

The term "time" may refer to a single instant or to a period. Typically, the statistical datum refers to an "instant" when it describes the "state" of a phenomenon (e.g. resident population on a given date). In this case we speak of "stock" data. The datum refers to a "period" when it describes "events" (e.g. variations in the resident population, such as births, deaths, immigration, emigration, during a period between two specified dates). In this case we speak of "flow" data.[26]

In terms of "time" and "groups of statistical units" that one wants to observe, two categories of statistical data are usually mentioned[27]:

- "cross-section" data, in which the phenomenon is observed at a single instant or in a single period of time and for many groups of statistical units (as in the preceding examples);
- "time series", in which the phenomenon is observed for a single group of statistical units and many instants or periods of time (generally at regular intervals).

Let us bear in mind, in any event, that as a rule one wants to observe many groups of statistical units with reference to many instants (or periods) or time.

[26] See [44], p. 5.
[27] See [50], p. 4.

1.2 Statistical data and analytical data

The peculiarity of statistical data is that they provide information on "groups" (sets) of objects[28] (e.g. the income of groups of people).

The analytical datum, by contrast, is generally considered as representing the characteristics of single objects (the income of individual households, the balance of individual bank accounts, etc.).

However, the same object may appear as "elementary" in one context or as a "set" in a different context.[29] For instance, a "household" may be considered as a group of persons, a bank account as a group of transaction flows.

It follows that the same datum (e.g. household incomes, bank account balances) may appear as analytical or statistical depending on how the reality under observation is conceptualized. For example, it is possible to consider "households" and "persons" as a different kind of elementary object, bound by specific relationships (each person belongs to a household), but it is also possible to consider them as sets of objects (the household as a set of persons and the individual person as a set composed of a single element, the person himself).

Hence there is no criterion for distinguishing between statistical and analytical data on the basis of "structural" characteristics. In practice, it will be convenient to use one or the other concept depending on the use one intends to make of the datum.

When data are used in support of "operative" processes, their structure is virtually unchanging, and any levels of grouping are generally preset and stable over time. In this case, it is generally convenient to adopt an "analytical" context.

When the data serve mainly to support processes of "inquiry", which are much less stable and predictable than "operative" processes, their structure and grouping levels can vary frequently with the needs of the inquiry, so a "statistical" context is more suitable.

Anyway, nothing prevents one from including in a "statistical" context data deriving from an "analytical" context, hence referring to objects originally conceived as elementary. In the statistical context, in fact, it is also possible to consider sets composed of single objects. A generic statistical datum, therefore, can always refer to a set, which in the most general case is composed of many elements but in specific cases may be composed of a single element (the object to which the datum refers).

Data relating to single objects are actually gathered as part of statistical data collection.

These are "microdata,"[30] i.e. data on individual statistical units from which, by the processes of aggregation and estimation, we obtain "macrodata", i.e. data relating to groups of statistical units.

Both the microdata and the macrodata, in the present work, are considered as statistical data. The idea is to find a common form of representation for them.

---

[28] The term "object" is utilized here in its intuitive meaning, as an element of reality having independent existence. The object may also be composite, as a vector of other objects (see [50], p. 2).
[29] See also [17].
[30] See [44], p. 2; [50], p. 2 ff.

1.3 Statistical functions

Returning to the argument of Section 1.1 and considering "time" as well, we can characterize the notion of "statistical datum" more precisely and more formally.

Suppose that we have defined a set of statistical units and a grouping,[31] a characteristic relating to the groups of statistical units (a statistical characteristic), a set of modes in which the latter can occur, and a set of modes of time in reference to which we are interested in observing the characteristic: the statistical datum is the law which, for each pair constituted by a group of statistical units and a mode of time, associates the mode of the statistical characteristic that occurs in correspondence with the pair.

The foregoing definition is a special case of the algebraic function,[32] which can be called the "statistical function".[33] The domain consists of a subset of the Cartesian product effected between the groups of statistical units and the modes of time, the co-domain consists of the modes of the characteristic. The algebraic notion of "function", therefore, enables us to formalize the notion of "datum" proper to descriptive statistics.

Like all functions, the statistical function can be described in two forms:

- intensional (structure), which corresponds to the "definition" of the function[34];
- extensional, which consists in the full expression of the law of correspondence between the elements of the domain and the co-domain.[35]

What follows is an example of how the intensional form of the statistical function "Italian household income" could be informally described.

function name:	**RFI**
description:	**income of Italian households**
statistical units:	**Italian households**
set of groups:	all **groups of households**
statistical characteristic:	**income** (billions of lire);
admissible modes:	**positive integers**;
time:	**years 1996, 1997**

[31] The grouping is a set of groups of statistical units, i.e. a sub-set of the power set of the set of statistical units.

[32] A function or application is a law that makes each element of a set D (the domain) correspond to one and only one element of a set C (the co-domain).

[33] The term "statistical function" is not used as in [33], pp. 6-7, but as a synonym for algebraic-mathematical function (the repetition of a given statistical experiment give rise to two distinct functions).

[34] The function is assigned a name, its meaning is specified, the domain and the co-domain are indicated.

[35] Whereas the extensional form of mathematical functions can also be represented concisely by symbols and combinations of symbols whose meaning is predefined (x^2, x*log(x), sin(x), etc.), this is impossible for statistical functions, which are derived empirically through observations and possibly subsequent processing. The form of statistical functions must thus be written out in full, by listing all the associations between the elements of the domain and the corresponding elements of the co-domain.

There is a corresponding extensional form like this:

GROUP OF HOUSEHOLDS	TIME		INCOME
A	1996	→	80
B	1996	→	171
…	…		…
A	1997	→	168
…	…		…

The domain and the co-domain are generally multidimensional.[36]

In the above example, each group is characterized by a "name" (A, B, …). As a rule, however, the grouping sets are identified by combinations of modes with appropriate systematic characteristics (Section 1.1).

For example, if we are interested in groups of households resident in a given region and having the same number of members, we may define a function "RFI-RN" (RFI by region of residence and number of members), associating with RFI (household income) as systematic characteristics, "region of residence" and "number of members of the household".

The grouping consists of groups of families residing in the same region and having the same number of members; it is represented by a two-dimensional space (region of residence and number of members). The domain is three-dimensional (the two dimensions that identify the grouping, plus time); the co-domain, which is one-dimensional, is the set of admissible values for the characteristic "income". The extensional form of the function is the law that specifies the income that corresponds to every combination of modes of region of residence, number of members, and period of time. For example:

GROUP OF FAMILIES		TIME	INCOME
Region of residence	**Number of members**		
Val d'Aosta	1	1996	71
Val d'Aosta	2	1996	83
Val d'Aosta	3	1996	128
…			
Piedmont	1	1996	68
Piedmont	2	1996	168
…			
Lombardy	1	1996	103
…			
Val d'Aosta	1	1997	75
…			

[36] For the co-domain, this occurs in the presence of a multiplicity of statistical characteristics or of multiple statistical characteristics (see [35], p. 29).

The intensional form of this function could be described as follows:

function name:	**RFI-RN**
description:	**income of Italian households by region of residence and number of members**
statistical units:	**Italian households**
set of groups:	all **groups of households** resident in the **same region** and having the **same number of members**
statistical characteristic:	**income** (billions of lire), with as admissible modes **positive integers;**
systematic characteristics:	**region of residence** (all Italian regions), **number of members** (integers from 1 to 99)
time:	**years 1996, 1997**

The components of the intensional form of the "statistical function" are thus the following:

- the set of statistical units;
- the grouping of statistical units into subsets;
- the characteristics of interest;
- the modes of the characteristics;
- time.

Hereinafter, the *set of statistical units* will also be called the "statistical class". This term is introduced in order to distinguish the set (the statistical class) from the elements of which it is composed (the statistical units) and is used to identify the set to which the data effectively refer.[37] Statistically speaking, this may correspond to the entire "population" or to a "sample" drawn from it.[38]

The composition of the statistical class in terms of statistical units can vary with time (for example, if the statistical class consists of "resident individuals", its composition changes with births, deaths, immigration and emigration).

The *grouping* is obtained by defining the sets of statistical units that we are interested in observing. The resulting sets must be subsets of the "statistical class", not necessarily disjointed, but such that their union coincides with the "statistical class" itself. If this condition were not satisfied, there would be statistical units not belonging to any subset: these would be neglected in the course of the statistical inquiry, and the proposition that we are dealing with statistical units would be negated.

[37] The statistical class corresponds to the "alpha" component of B. Sundgren's "box structure" (see [44], pp. 4-5, and [47], pp. 10-11), cited in the proceedings of the Conference of European Statisticians (see [50], p. 4); in the MCS model, it corresponds to the "class of objects" (see [17], p. 408).

[38] Depending on whether the datum refers to the whole population or to a sample (see [33], pp. 9-10).

In the most general case, the grouping is constituted by all possible subsets of the statistical class (given a set of "n" elements, there are 2^n possible subsets). For instance, if the statistical class is the set {Luke, Mary, Joe}, there are 8 possible subsets:

subset 1: {Luke, Mary, Joe}
subset 2: {Luke, Mary }
subset 3: {Luke, Joe}
subset 4: {Mary, Joe}
subset 5: {Luke }
subset 6: {Mary }
subset 7: {Joe}
subset 8: { }

In this example, we have used the extensional form for defining a set,[39] which can obviously also be used in the case in which we are not interested in defining a grouping that comprises all the possible subsets of a statistical class.

When the sets of groupings are defined by means of one or more characteristics of the statistical class (which is a very powerful and general method of definition), by contrast, one uses the intensional form: that is, one enunciates a property that the elements must satisfy. Every combination of modes of the characteristics identifies a set, whose members are the statistical units for which that combination occurs.[40] The number of sets defined will depend on the number of characteristics we are interested in and how many modes they have.

Both the sets of the grouping and their composition may vary with time.

The characteristics describe the results of the inquiry (statistical characteristics) or else define the grouping sets (systematic characteristics).[41]

A characteristic is a property type (e.g. age is the time that has passed since birth) that can occur in a predetermined set of "modes" (the modes of age, for example, could be "years of age") and to which there is associated a procedure which, given a generic mode, permits us to determine whether it occurs or not in correspondence to each group of statistical units for which that characteristic has sense (and with reference to a given "time").

A characteristic is thus fundamentally a set (the set of the modes in which it can occur) associated with a precise meaning (the meaning with which the modes of the characteristic are interpreted).[42] Therefore one must consider characteristics to be different if they have different meanings, even if they occur in the same modes (e.g. "working age", defined as the time passed since one's first employment, is different from civic registry age, even though both may occur in the set of years).

[39] This consists in listing the elements of the set explicitly.

[40] Grouping by means of characteristics corresponds to the "gamma" component of B. Sundgren's "box structure" (see [44], pp. 4-5, [47], pp. 10-11, [50], p. 4) and to the "statistical classification" of the MCS model (see [17], p. 409).

[41] Systematic characteristics correspond to the "gamma variables" of the "gamma" component of Sundgren's "box structure" and to the "category attributes" of the MCS model; the statistical characteristics correspond to the "beta variables" of the "beta" component of the "box structure" and to the "datum classes" of the MCS model (see [44], pp. 4-5, and [47], pp. 10-11; [50], p. 4; [17], pp. 408-09).

[42] The notion of "characteristic" presented here corresponds to Sundgren's "variable" rather than to the "object characteristics" or "statistical characteristics". The latter, in fact, are associations between "objects" or "sets of objects" and the corresponding "variables" (see [47], p. 23 and [50], pp. 2, 4).

The attribute "statistical" or "systematic" refers to the role (as dependent or independent variable, respectively) that the characteristic takes on within the statistical function, not an absolute feature of the characteristic. In principle, there is no reason why a characteristic cannot be considered as "statistical" for one function and "systematic" for another.

The *modes* of a characteristic are freely predetermined as a function of what we are inquiring into. They depend in particular on the level of detail desired. For instance, when age is under scrutiny, we may be interested in days, years, or broader age-classes (youth, adult, elderly).

Consistent with the statistical context, a mode can generally be considered as representing a set of objects (which may be composed of many objects or of only one).

In the case of the characteristic "individuals' age-classes", say, the modes correspond to a set of years (e.g. youth=0-18, adult=19-60, elderly=over 60). Similar considerations can be made for other qualitative characteristics obtained by subdividing the field of definition of a quantitative characteristic into intervals.[43] In practice, moreover, any continuous quantitative characteristic is turned into a discrete series of values, owing to the finite precision of measurement. Every mode thus actually identifies an interval of values.[44]

The modes of qualitative characteristics can also be considered as representing sets of objects. This is obvious when the modes express hierarchically differing levels (e.g. continents, countries, and cities as modes of "geographical location").

It is worth repeating that the modes in which a characteristic occurs are not necessarily disjointed.[45] This is the case, for instance, of hierarchically differing levels of detail.[46]

The sets of modes in which characteristics occur can also vary over time.

Time, finally, is an essential component, because as we have seen all other components exist and evolve in relation to time (e.g. statistical class, groups, characteristics, etc.).

Actually, there can be more than one time parameter of a function.[47] In the present work, however, for simplicity we shall consider only one "reference time".[48]

[43] See [33], p. 14.
[44] See [33], p. 15.
[45] This concept is meaningful when the modes are considered as "sets".
[46] See the previous example and also [44], p. 6 (hierarchical gamma variables).
[47] See [47], p. 8-11 and [50], p.4.
[48] "Reference time" corresponds to the "tau" component of Sundgren's "box structure" (see [44], p. 5; [47], pp. 10-11; [50], p. 4).

1.4 Statistical concepts

Statistical data are normally defined with concepts.

For example, in the RFI-RN function (see section 1.3):

- the statistical class refers to the concept of **"Italian household"**;
- the statistical characteristic refers to the concept of "**income"**;
- the systematic characteristics refer respectively to the concepts of "**region of residence**" and "**number of members**";
- the modes of the systematic characteristics refer respectively to the individual regions (**Piedmont, Val d'Aosta, etc.**) and individual numbers (**1, 2, 3, …**).

A concept expresses a given meaning, i.e. the result of an abstractive process, which essentially consists in isolating certain properties from reality as a whole, ignoring others that are not considered relevant in the given circumstances. The concept is the unifying element of the properties of interest.[49]

"Statistical concept" will be used to identify all concepts, in the sense described above, that are of interest in statistical enquiry, especially those used to represent statistical data.

Thus, a statistical concept is an abstraction that defines an aspect of reality that is of use for statistical purposes.

It is assumed that a concept is independent, i.e. that it can exist independently of the fact that one, none or many "statistical data" fall within its definition.

By contrast, a statistical datum refers to many concepts and cannot exist if these do not exist.

A single concept can serve as a reference for many statistical data, but its meaning does not depend on these data. For example, whatever datum refers to the concept of "income", such as:

RFI	Italian household income;
RFI-N	Household income by number of members;
RSI	Italian corporate income;
….	

the meaning of "income" remains the same.

In referring to a concept, a statistical datum also specifies the role in which the concept is used, i.e. whether it is what is being described (statistical class, groups of statistical units, systematic characteristics) or what describes (statistical characteristic).

[49] See also [1].

The same concept can be used by different data in different roles. For example, take the following statistical data, which refer to the same concept ("country").

- income of inhabitants grouped by **country** of residence;
- number of **countries** grouped by form of government;
- main destination **country** for exports grouped by category of good.

In the first case, the concept "country" falls under the definition of the systematic characteristic; in the second it is a statistical class and in the third it is used to describe a property type, i.e. in order to describe a statistical characteristic. It is clear that the meaning of "country" is the same in all cases; what changes is the role that the concept plays within the statistical function.

1.5 Metadata and metainformation

In the literature on information systems, "datum" refers to the physical representation of "information". It then uses the term "metadata" to indicate data regarding other data and "metainformation" to indicate information describing other information.[50]

The extensional form of statistical data can be considered a "datum", from which it is possible to draw certain information regarding the reality under observation, while the intensional form of data and the definition of concepts are metadata that provide metainformation.

We then have "information systems" that process information and "metainformation systems" that process metainformation regarding an information system. The attribute "statistical" is added if the systems handle statistical information and the related metainformation. An information system describes the "objects" of reality, while a metainformation system describes the objects of the information system (metaobjects).

The primary objective of an information system is to construct a "complete and self-consistent" representation, i.e. one that in addition to containing data (information), also contains the metadata (metainformation) needed to understand the meaning of the data. Such a system is "infologically" complete.[51] The advantage is clear: the system contains all the information needed to interpret the data.

In order to meet this requirement, a statistical information system must, in addition to the extensional form of the data, also include their intensional definition and the definition of the concepts.[52]

An equally important objective is to supplement the information resources with the information necessary to administer and manage the system for all categories of users (end users, administrators and software).[53] Such a system can be called "procedurally complete".[54]

In addition to "conceptual" metainformation, procedural completeness requires the availability of metainformation linked to the specific demands of the implementation,[55] which are not considered here.

Procedural completeness also implies that metainformation (conceptual or otherwise) is accessible to people as well as software. This is ensured by storing data and the related metadata within the same database.[56]

Procedural completeness must be accompanied by "intrinsic congruence": the operation of the information system must ensure that the information and metainformation are always consistent.

This requires the "active" use of metadata as "rules" for data processing.[57] First, the software must be "generalized", i.e. able to execute its functions on any statistical datum whose intensional

[50] See [45] p. 3 ff.
[51] See [45] p. 12.
[52] This are considered "conceptual" or "infological" metainformation.
[53] See [45] pp. 9-10.
[54] See [45] p. 12.
[55] Logical and physical levels of the representation (see [16] pp. 125-126).
[56] See the approach taken by Magnani in [29].
[57] For example, one can use the intensional form of the datum to "control" the corresponding extensional form.

form is represented in the system, using the metainformation as a guide for its processing.[58] At the same time, in order to ensure that the metainformation can be easily used by the software, it must be represented with formal rules.

Although "formal" representations avoid the ambiguities inherent in natural language, they cannot replace it entirely. Natural language links the metainformation system with external reality and is the means used to express all the information that cannot or should not be formalized (for example, descriptions or methodological notes).

From the point of view of "content", information systems (and the related metainformation systems) obviously differ from case to case.

What is of interest here, however, is not the individual case but rather the general structures and rules for the representation of the intensional form of data and statistical concepts (i.e. the model). The application of a model to specific cases enables the production of specific representations, each concerning an individual case.[59]

A model is all the more general the greater is its capacity for representation and the larger is the category of problems that it can solve. This paper is primarily concerned with the issues regarding the automation of a statistical information system[60] supporting a variety of activities, such as:[61]

- the design of statistical surveys;
- the performance of statistical surveys (collection);
- the coordination of different surveys;
- the storage of statistical data in a database open to multiple users, containing the results of a variety of surveys and information from numerous sources;
- the production of processed data (aggregates, estimates, etc.) and their storage in the database;
- the use of the statistical information system by end-users;
- the design, production and systematic distribution of statistical products.

As a result, our attention is focused on the formalization of the metainformation that will play an active role in processing.

[58] For example, a generalized aggregation function must be able to produce a generic statistical datum regarding the European Union from the corresponding data for the member states, basing its operation on the intensional forms and the relationship between the European Union and the member states (represented in the metadata system).
[59] The literature sometimes uses "model" for the individual representation and "metamodel" for models.
[60] In which the metadata contained in the schema play an active role, i.e. guiding processing.
[61] See also [45] p. 4.

1.6 The macrostructure of the representation

A model for describing and administering data and statistical concepts should enable the representation of the types of information and metainformation described in the previous sections. The decomposition of this complex problem into simpler sub-problems leads to the identification of three categories.

The first concerns the abstract definitions, i.e. the statistical concepts. It is the prerequisite for the definition and representation of the statistical data.

The evolution of the concepts is relatively independent of that of the data:

- concepts are definable whether or not any data refer to them;
- statistical data can be fully defined only with reference to previously defined concepts;
- a concept can be used in the definition of different data and in different roles;
- the meaning of a concept does not change in relation to the data that refer to it.

The different nature and independent evolution of statistical concepts justifies separating their representation from that of the statistical data and using representational structures appropriate for this specific case. A first component, called the "statistical concepts dictionary", is therefore incorporated into the architecture of the information system. It contains the definition of all the concepts of interest. The expression "administration of statistical concepts" means populating and updating the concept dictionary.

Let us now turn to *statistical data*, or statistical functions. For the purposes of representation, a distinction must be made between the *intensional form* and the *extensional form*. In a first approximation, the intensional form specifies the datum even before it is observed, while the extensional form describes the result of the observation.

Here, too, it is useful to use different representational structures owing to the different nature and dynamics of the two categories:

- the "intensional" form of a datum corresponds to the algebraic definition of a function, while the "extensional" form is the extended representation of the function itself;
- the definition of a function may be available regardless of any knowledge of the extensional form (in statistical surveys this is often the case, if for no other reason than because it is usually necessary to determine "what" we wish to observe before actually observing it).

The job of containing the definition of the intensional form of the data belongs to another component of the system architecture, know as the "statistical data dictionary". It includes the statistical functions and the related information, such as the procedures used to obtain that information, the reciprocal relations, etc.

These components, that is the concept dictionary and the data dictionary, are called the "statistical dictionary". This paper focuses on this aspect. The third architectural component, which contains the extensional form of the data, is called the "statistical data base". The expression

"administration of statistical data" refers to the activity of populating and updating the statistical dictionary and data base.

This decomposition helps to rationalize representations and simplify administration. It is based on elementary principles of normalizing representations.[62] For example, since multiple data can refer to the same concept, the representation of the properties of the concepts and the reciprocal relationships within the concept dictionary minimize the work needed to produce definitions and lower the risk of inconsistency.

However, the three components must contribute to describe a datum fully, analogously to the description of a "sentence" of natural language:

- the meaning of the terms must be in the dictionary;
- the syntax describes the structure of the sentence;
- the sentence itself is represented on a sheet of paper, in a book or on some other medium.

The components are bound together by reciprocal links and constraints that ensure the integrity of the representation. The three components also contain the accessory metainformation, such as methodological notes, non-key metadata[63] and any other information required, depending on the category to which it refers (concepts, intensional forms, extensional forms).

[62] See [16] pp. 142-144.
[63] See [38].

2. STATISTICAL CONCEPTS

2.1 General

Statistical concepts are "abstractions" of the aspect of reality that we are interested in.[64]

Such concepts are formed (more or less explicitly and deliberately) at the very act of conceiving a statistical enquiry. In fact, concepts are used to design the observations, i.e. to define the "intensional form" of the data.[65]

Concepts are also a useful guide in producing the "extensional form" of the statistical data (the final result of the survey). For example, they are used (more or less deliberately, either automatically or manually) to specify the data to be gathered, to check the data collected, to determine the processing algorithms and so on.

Interpreting the data obtained is also based on understanding the "concepts" to which the data refer.

Concepts define all elements that can be used in making observations, such as the groups (e.g. individuals), their properties (e.g. "income", "marital status", "country of residence") and the modes in which such properties may be expressed. They are the building blocks in the design and representation of statistical data.

[64] See section 1.4.

[65] For example, if we wish to measure the income of individuals in relation to country of residence and marital status, it would be necessary to have defined (or at least conceived of) the concepts of "income", "individual", "country", "residence" and "marital status".

2.2 Concepts and the field of observation

Statistical studies are carried out in order to learn about a given segment of reality, ideally abstracted or separated from the rest of reality, which is known as the "system of interest"[66] or, in this paper, the "field of observation" (for example, demographic phenomena, labour relationships, credit and finance).

The mental model of this segment of reality consists of a complex of aspects "of interest", i.e. statistical concepts.

Concepts have two basic properties: existence and meaning.

Existence is linked to our interest in considering a certain aspect of reality. There is an infinite number of concepts that can be conceived, but only "useful" concepts are considered to "exist".

Meaning may initially be expressed in natural language (definitions, descriptive notes, etc.).
It emerges from the area of human knowledge at which the field of observation is directed (for example, demographics or economics) and is the "link" between the field of observation and external reality.

In addition to concepts, the relations between the concepts may also be important in the field of observation. For example, there is a specific relationship between the concept "Europe" and the concepts for the individual countries: Europe is composed of the countries Italy, France, etc.

Relations also have existence and meaning (in fact, relations can be considered to be concepts), and again the existence of a relation depends on the interest that relation holds (only relations useful for statistical purposes are considered to exist).

It can be very helpful to represent concepts and relations in formal terms because this allows us to improve the specification of the meaning of the concepts. Natural language may not be sufficient to avoid ambiguity. Take, for example, the concepts "region of residence" and "region of domicile". It can be difficult to deduce from their respective natural language definitions whether the two are synonyms or two different concepts. This deductive process may produce different answers depending on who is doing the deducing. The formal representation of a relationship of "synonymity" (concepts with equivalent meanings) can eliminate this ambiguity.

Formal expression also facilitates the use of concepts and relations in automated processing.

Relations between concepts, for example, can be "inherited" by the data to which the concepts refer and used for processing the data (e.g. the relation between Europe and the member states can be used to calculate the number of people or the income of Europe starting from the corresponding data for the individual countries).

Knowledge of concepts and their relations is essential to compare the data on which the concepts are based (for example, understanding the relation between Europe and the individual countries makes it possible to correlate the related data).

[66] See [45] p. 3-4.

Precisely because it facilitates comparability, the field of observation is broadened and enriched with the integration of statistical surveys. For example, if it is necessary to use data from different surveys together, it is also necessary to correlate the related concepts, i.e. consider them as part of the same field of observation.

In conclusion, the field of observation is a coordinated complex of definitions for concepts and relations between concepts that describes the abstraction of a segment of reality and that are a prerequisite for the definition and understanding of the statistical data that refer to that segment. In particular, the field of observation of a statistical information system contains the relevant concepts for all the data that the system contains.[67]

[67] It is represented in the concept dictionary (see section 1.6).

2.3 The role of time in the field of observation

All aspects of reality are intuitively perceived in a temporal dimension.[68]

On the one hand, time is a "primitive" statistical concept (i.e. it exists in all fields of observation and is not defined in terms of other concepts). On the other, all other concepts (or relations between concepts) are linked to a specific moment or period of time.

The field of observation is therefore intrinsically "historic", that is it describes the existence of concepts in relation to time.[69]

Since a concept exists because we are interested in it (see section 2.2), the existence of the concept in the field of observation may also not coincide with the actual existence of the "objects" of which the concept is an abstraction. For example, even if in reality "countries" have existed for hundreds or thousands of years, the concept of "country" is considered to exist only in relation to the period of time for which it is of statistical interest.[70] Conversely, although the country "East Germany" no longer exists as a political reality, it could still be of interest in statistical reality.[71]

For simplicity, we follow the intuitive conceptualization in which the identity of a concept does not depend on time and the concept itself exists in relation to time.

A less-intuitive approach consists in basing the conceptualization on the "object-moment" pairs.[72] In this case, a concept corresponds to the abstraction of one or more objects of reality referred to a specific moment in time.

[68] In all the cases of interest in this context.
[69] They are born, exist and die in time.
[70] Which is generally the period for which we are interested in the statistical data that refer to the concept.
[71] The consideration of "time" and the "existence" of concepts can be compared to the statement of infological models (see the summary in [48] pp. 226-242).
[72] See [44] p. 5, "the tau dimension".

2.4 The formalization of statistical concepts

Statistical concepts can be divided into categories corresponding to types of components that make up the definition of a generic statistical datum, i.e. statistical classes, characteristics, sets of modes, time (see section 1.3).

In order to formalize the representation, earlier we identified an algebraic correspondent of the notion of "datum", i.e. the "function" (see section 1.3). Similarly, each category of concepts can be associated with one of the notions that make up the definition of a generic function in algebra, i.e. "variable", "set" and "element" of a set.

In this way, characteristics correspond to the "variables" of a statistical function, sets of admissible modes of characteristics correspond to the "sets" in which the variables have values, and the individual modes are the "elements" of those sets.

In turn, time can be regarded as a special variable: "reference time".[73]

The grouping of statistical units can be represented by means of variables, for example by a "variable" that assumes values in a "set" whose "elements" correspond to the groups: each "element" therefore identifies a group of objects in reality.[74] The grouping can also be represented by means of variables corresponding to the "systematic" characteristics.

The variables (and their respective sets) exist in relation to the interest that they have in a specific field of observation. The existence of a time variable (and the respective set in which it assumes values) is expected to be essential in any field of observation.

As a result of their time dependence (see section 2.3), "variables", "sets" and "elements" of a field of observation are generally "historical". This property distinguishes the categories of "statistical" concepts from their "algebraic" namesakes.

A "historical" concept (for example, a historical set), considered at a specific moment, corresponds to its equivalent "algebraic" object (for example, an algebraic set), and, throughout its existence, can be considered as a collection of algebraic objects, each referring to a specific moment.

[73] It is the variable to which the existence of statistical concepts is referred and which expresses the time in relation to which a group of statistical units possesses a certain property (see section 1.3).

[74] Such a variable is related to the "statistical class", in fact its elements correspond to groups of statistical units.

2.5 Statistical variables

"Statistical variables" are the algebraic transposition of the notion of "characteristic".[75] They represent the different meanings with which sets can be used to describe reality.[76]

For example, the set of "countries" can be used to describe the residence (country of residence) or birth (country of birth) of an element of another set (people, companies, etc.).

A variable (for example, country of residence) therefore represents a specific meaning with which a set of elements (countries) can be used to describe a type of property that can be referred to another set of elements.[77]

The existence of a variable (country of residence) presupposes the existence of the set in which it appears (countries). In this sense, variables do not have an independent existence.[78] The relation between a variable and the set in which it appears can be summarized as follows: each variable (country of residence) takes on a value in only one set (countries), but the same set can give values to more than one variable (country of residence, country of birth, etc.). In other words, the same set can have different interpretations.

A variable (for example, "income") can be used as a descriptor for many sets (people, companies, etc.), provided that the type of property makes sense for them (for example, the variable "income" has meaning for the set of persons or companies but the variable "gender" does not have meaning for the set of companies). A variable exists independently of the existence of the sets of elements that can be described by the type of property that it represents.[79] The "country of residence" can exist independently of the sets "persons", "companies", etc., whose elements reside in a country.

[75] The term "variable" is used to indicate both quantitative and qualitative characteristics according to the mathematical meaning, while sometimes in the literature different names are used to distinguish them (for example see [35] pp.28-29).
[76] See also the notion of "attribute" in [48] p. 24.
[77] In reality, it can also be used as a descriptor for a property of elements of the same set. For example, the "partner country" in international trade.
[78] Note that the existence of a statistical variable, like any other concept, is related to time (see section 2.4).
[79] It may nevertheless have no practical value to define variables that do not describe any set.

2.6 Statistical sets

The set in which a variable has a value corresponds to the set of the admissible modes of the characteristic.

Sets are conceived as concepts with independent existence (countries, age-classes, etc.),[80] composed of elements that also exist independently (individual countries, individual age-classes, etc.). They can be considered to exist independently of the existence of variables that refer to them.[81]

Sets can contain elements that correspond to material objects (persons, automobiles, places, etc.), mental abstractions such as events or relations (marriages, sales, etc.),[82] to modes of manifesting properties, including those for which we perceive an "objective" existence (age, genders, incomes, durations, etc.) and those that are the result of abstraction, such as categories, classes, judgements (the categories of economic activity, age-classes, income classes, quality levels, etc.).

A statistical set is defined using both of the classic methods: the specification of the properties that its elements must possess (the intension of the set) and the list of the elements of the set (the extension of the set).

The "meaning" of the set corresponds to its intension. The specification of the "properties" of the elements is comparable to the definition of a "concept" as an abstraction of the common properties.[83] The intension may be specified with a statement that qualifies the properties of the set, defining its "potential" elements (for example, a "country" is the set of points of a territory subject to the same jurisdiction).

The extension is a list of the "actual" elements of the set (existing countries). In some instances it can also be expressed in summary form. For example, "numerical" sets can be specified with an interval of values, e.g. integers from 0 to 10^{15}.

The extension of a set generally changes over time. The set of countries, for example, changes its extension whenever countries are created or disappear.

Statistical sets can be "composed" with the usual set operators (union, intersection, classification, Cartesian product, power set, etc.), which are applied to a specified time period to obtain other statistical sets. The intensional form of the resulting set can be obtained from the logical composition of the statements that define the "operand" sets, while the extensional form can be obtained by applying the set operator, moment by moment, to the algebraic sets that correspond to the "operand" sets in the moment considered.

[80] Note that the existence of a statistical set, like any other concept, is related to time.
[81] It may nevertheless have no practical value to define a set to which no variables refer.
[82] See [50] p. 2.
[83] See section 1.4.

2.7 The elements of statistical sets

The "elements" of a statistical set are the algebraic transposition of the notion of "mode" found in descriptive statistics. They exist independently.[84]

Statistics makes a specific assumption in considering "modes" (i.e. elements) as in turn representing sets (see section 1.3).[85]

One might wonder what difference there is between modes and sets of modes (and thus between elements and sets of elements), given that both are considered sets. The answer is that while modes correspond to sets of objects that are not formally represented in the field of observation (for example, "countries" are sets of "points of the territory", which are not explicitly represented), sets of modes are sets of objects that belong to the field of observation. We therefore also want to represent the extension (list of components) for sets of modes, while the intension (characteristic property) is sufficient for modes.

Although the extension of the "elements" is not represented, it does exist and can change over time. For example, "persons resident in Italy" (a hypothetical element of the "statistical class" set) change over time as a result of births, deaths, immigration and emigration; "France" (an element of the set of countries) can change over time as a result of acquisition or loss of territory.

Elements corresponding to manifestations of time can also be conceived in the same manner: a period of time is considered as a set of moments; a single moment is a set containing just one moment.

As they are considered sets, elements can be "composed" using the set operators (union, intersection, partition). For example, "Italy" together with "France", "Spain", etc., is equal to "Europe". However, this does not mean that the result of the composition belongs to the same set to which the operands belong. This will depend on how the set is defined. In the example, if the set to which the operands belong were "countries and continents", Europe would belong to that set, while if the set were "countries" Europe would not belong, since it is not a country.

[84] Note that the existence of an element, like any other concept, is related to time.
[85] This assumption produces a special algebraic structure, which we will discuss later.

2.8 Attributes, relations and integrity constraints

The basic properties that distinguish variables, sets and elements of sets are two: existence and meaning.

Existence is time-dependent and can be qualified by one or more temporal attributes (e.g. time of birth and death).

Meaning can be expressed, in the first instance, through a defining statement (see §2.2, 2.6).

Depending on the needs of a specific field of observation, it may be of interest to describe other types of property[86] (attributes). These comprise various forms of "metainformation". In the case of variables, for example, it may be useful to specify whether they are qualitative or quantitative, whether they express a raw or a seasonally adjusted value, a stock or a flow, their physical dimensions (cardinality, length, volume, mass, monetary value, …) or statistical indicators (sum, average, variance, …) etc.

Between "variables", "sets" and "elements" there exist types of relationship[87] that induce specific integrity constraints.[88]

As a consequence of the time-dependence of statistical objects, the relationships between them are also time-dependent[89] and integrity constraints must be satisfied in reference to every instant.

A first type of relationship is established between variables and sets; it links the generic variable to the set in which it may assume values. Each variable, in each instant of its existence, assumes values in one – and only one – set. A set, on the other hand, can provide values for more than one variable (for example, the set of "nations" gives values to the variables "nation of birth", "nation of residence", "nation of export", etc.). The set in which the variable assumes values can also be referred to as the variable's "domain of definition".[90]

A second type of relationship is established between sets and elements; it links the generic set to its constituent elements.

At any instant of its existence, a set can contain any number of elements that exist in that instant (or it can even contain no elements, in the extreme case of an empty set). An element, at any instant of its existence, can belong to numerous sets. For example, the element "Italy" may belong to the set of the nations of the world, of the nations of Europe, of the member-nations of NATO, of the territorial areas of Italy, and so on.

---

[86] There is a distinction between a single "property" belonging to a single element (e.g. "France" is the nation in which "Giovanni" resides) and a "type of property" (a set of single properties) that characterizes a set of elements (e.g. the "nation of residence" of "persons").

[87] Here, too, a distinction is made between a single "relationship" linking specific objects and a "type of relationship" that is established between types of objects.

[88] For the definition of integrity constraints, see also (16), pp. 153-155.

[89] In other words, they express links that exist in relation to time.

[90] The following chapter describes the situations in which a variable assumes values in a sub-set of its domain of definition (see §3.7).

The types of relationship just described are essential in every field of observation. Other types may be of interest, depending on the informational requirements of the specific field involved.

For example, it may be useful to represent the equivalence of meaning (synonymity) between variables, or between sets or elements. With regard to variables, moreover, it may be useful to describe the composition of multiple variables in terms of simple variables.

It is particularly useful to describe the set relationships that are established between sets or between elements of sets.

Set operations (as in §2.6 and 2.7 above) induce corresponding relationships between operands and the result.

One example could be the type of relationship "is the intersection of", between statistical sets, one case of which might be: the set of the "nations of Europe and of NATO" is the intersection of the set of "nations of Europe" and that of the "nations of NATO".

Again, the type of relationship "is the union of" coming for example between "elements", one example of which might be: the element "European Union" is the union of the elements "Italy", "France", and so forth.

The special relevance of inter-set relationships is further clarified in the following paragraphs.

2.9 Levels of detail

One of the characteristics of statistics is the multitude of levels of detail on which we can view reality. These levels, moreover, change over time not only in relation to the evolution of reality, but even more so in relation to the changing focus of the observer, who of the two is undoubtedly the more dynamic.

An "element" of a statistical set, because it is itself equivalent to a set, corresponds to a specific level of detail in the description of a particular aspect of reality (known as "domain"). For example, in terms of the increasing detail used to describe a territory, the element "Europe" is at the level of continent, the element "Italy" is at the level of nation, and the element "Rome" is at the level of city.

More abstract notions such as categories of economic activity, for example, are also commonly conceptualized at numerous levels of detail[91] (starting from the "elementary" types of economic activity, "compound" types can be derived at various levels of composition).

There may also be "isolated" elements (i.e. elements that do not belong to any partition of a specific level of detail), such as "European Union", "eastern Europe", "NATO", "ONU", "South-East Asia", etc. which are neither cities, nations or continents, but which are none the less areas of territory.

An operator that is of fundamental importance in statistics is that of "set union" of "elements", since it allows us to relate two different levels of detail (the higher level of the operands and the lower level of the result).[92] This union is recursive, since it allows us to define groups of groups on many levels (e.g. groups of "cities" form "nations" and groups of nations form "continents").

Of particular interest is the union of disjointed "elements" (as in the examples described), given that, as will become clearer afterwards, the union of disjointed elements often corresponds to the "sum" of the related data (e.g. the number of inhabitants of Europe is the sum of the numbers of inhabitants of the individual nations that comprise Europe). In this case we talk of "set aggregation". The opposite operation is partition (the breaking down of an "element" into "disjointed elements" which, when rejoined, again constitute the original element). This is called a "set disaggregation".

The union of disjointed "elements" and partition are two operators (from greater to lesser detail and vice versa) that induce the same type of relation between elements (the element "x" is the "union of", or is "divided into" the disjointed elements x_1, x_2….).

By uniting all the elements of a statistical set we obtain the element that corresponds to the minimum level of detail relative to the set: this element is known as the "universe" of the statistical set.[93] The elements of the set are thus subsets of the universe. The "complement" of an element in a set is the set difference between the universe and the element.

[91] See also [44] p.6 (hierarchical gamma variables)
[92] The resultant element may not belong to the same set as the operands (see 2.7).
[93] The "universe" element represents the set conceived as a whole (undivided) and may belong to the set or not.

There are two particularly important methods of defining the intension of a statistical set, which spawn two particular categories of sets.

The first consists in defining the elements of the statistical set as a specific set disaggregation of its universe. One example is the set of nations of the world, given that it corresponds to a specific breaking down of the world's territory into disjointed sets. This is known as a "partition set".

A set composed of different hierarchical levels (e.g. the set of cities, nations or continents) can also be considered as the union of different partition sets, one for each level.

The second method consists in defining the elements of the statistical set as elements of the power set of the universe.[94] One example of this is the set of all the areas of territory (each area is a subset of the whole territory). In this case we refer to the "set of the parts" (or "power"set).

If a particular universe is chosen, all its partitions are included in its power set. Therefore, the "partition" sets are subsets of the "power" set.[95]

There can, of course, also exist sets that are neither "partition" or "of the parts" and which correspond to intermediate situations. By examining the two categories mentioned, however, it is possible to acquire a general understanding of the algebraic structure and utility of the different types of set in relation to representing the levels of detail.

[94] The power set of a set is the set of all its sub-sets.
[95] This is also true if the universe of the "partition" set is a subset of the universe of the "power" set.

2.10 “Partition” sets

A “partition” set contains the elements to describe a certain aspect of reality down to a single, specific level of detail.

It is the type of set in which a statistical variable that is destined to classify a specific aspect of reality (the universe of the partition) assumes a value. For example the variables “nation of residence” and “nation of birth” assume values in the set of “nations” that is a partition of the universe (“world territory”).

Its elements, defined as a set disaggregation of the universe, are all disjointed. The set, naturally, does not contain its universe (except in the extreme case of sets of a single element). However one chooses two elements belonging to the set, their union, intersection or complement generally no longer belong to the set. Intersection always gives an empty set, union always gives elements that do not belong to the set, the complement gives the other element of the set only in the specific case of a set comprising only two elements.

The elements of “partition” sets may be part of an aggregation (or disaggregation) only with elements of other sets (to each level of detail there corresponds a different partition set). For example, each nation is the union of specific cities, each continent is the union of specific nations.

The addition or elimination of levels of detail involves the addition or elimination of the corresponding partition sets; their configuration thus varies according to the levels of detail.

2.11 "Power" sets

A "power" set (e.g. areas of a territory) may contain all the elements that describe a particular aspect of reality, at any level of detail. It may contain both elements derived from disaggregations of the universe – in other words, belonging to "partition" sets (such as nations, continents) – and "isolated" elements (such as the "territory of NATO nations", "territory of UN member nations", see §2.9).

The aggregation (or disaggregation) of the elements of the "power" set also produces other elements that can be defined in the set. Relationships of aggregation (or disaggregation) that correlate different levels of detail are thus established between elements of the same set.

The addition or elimination of levels of detail does not, as it does with "partition" sets, involve modifying the configuration of the sets, but only the addition or elimination of elements and of the relationships in which they participate. In a statistical context, in which the levels of detail may frequently vary, the "power" sets have the advantage of ensuring the invariance of the configuration of the sets in respect of the variation in the levels of detail. For the above reason, a "power" set is suitable for collecting the definition of the "elements" that describe the same aspect of reality at any level of detail.

The intensional definition of these sets satisfies the property of "closure" of the union, the intersection or the complement of elements[96] (the property is taken to be verified in respect of the intensional definition, i.e. the concepts potentially "definable" in the set).[97] For example, the set of "nations" (a "partition" set) is not closed in relation to the union; the union of EU member states supplies the element "European Union", which cannot belong to the set because it is not a nation. Vice versa, the set of the "areas of territory" (a "power" set) is closed in relation to the union, since the union of any couple of areas supplies another area (the EU is also an area of territory), which can belong to the set.

A "set of the parts" relative to the union, the intersection or the complement of elements, possesses a Boolean algebraic structure.

[96] The set "K" is closed to the operator "#" if, however the two elements k_1 and k_2 , belonging to "K" are chosen, also the element $k_3=k_1 \# k_2$ belongs to "K".

[97] The element k_3 must be definable in the set, i.e. it must satisfy its characteristic property.

2.12 Probabilistic analogy

The algebraic structure of a "power" set, in other words the algebra that is established between modalities at different levels of detail, is totally analogous to that of the algebra of the events relative to the axiomatic formulation of the theory of probability.[98]

This coincidence is not a casual one. A comparison between descriptive statistics and the theory of probability shows that every population can also be described using conceptual instruments proper to the theory of probability. There is in effect always the possibility of studying some feature of a population in terms of a random variable.[99] To this end a special chance phenomenon can be defined: using the urn model, this consists of the random extraction from an urn containing the population or sample to be studied and finding the modality shown for the element extracted.

A series of correspondences is thus established between notions of descriptive statistics and those of the theory of probability. With reference to a single instant, in fact:

- The "set of statistical units" corresponds to the contents of the urn (e.g. individual persons);
- a "character" corresponds to a random variable (e.g. age-classes);
- the set of admissible modalities of a character corresponds to the set of admissible events of the corresponding random variable (e.g. [young, mature, old]);
- the single modality corresponds to an event (e.g. "young").

The probabilistic structure of the random variable is described by the relative frequencies of the modality. In the present context, however, we are not interested in determining the probability that a given event will happen (in other words, a given modality), but rather in deriving a formal representation of the events, for which purpose the primitive notions, postulates and conclusions proper to the theory of probability can be used.

The present formulation, nonetheless, differs from the theory of probability in its consideration of time. A "probabilistic" event is not time-dependent, whereas "statistical" events exist and must change their configuration over time (in other words they are "historical") and must be considered as a collection of probabilistic events, each one in relation to a specific instant.

The theoretic path outlined in the above paragraphs could also have been developed by starting from the theory of probability. The following paragraphs will link up again with this theory, partly using notions and terms borrowed from it, to continue the discussion.

The considerations outlined so far thus remain valid even when the "variable" is taken as a "random variable", an element represents an event and a set of elements corresponds to a set of events.

[98] See [35] p. 119 ff.
[99] See [35] p. 261,262.

2.13 The space of the results

A "chance phenomenon" consists in implementing a "process" that produces a given "result". Before performing the process the result is unknown, but the set of possible results is.

For example, if the process consists in throwing a die, the possible results are represented by the six faces of the die. If instead it consists in extracting a person from an urn and observing the place of birth of the person thus extracted, the possible results are the points of the earth's surface.

In the vast majority of cases (all of those that are of interest in the present context) it is possible to associate each result with a point in a Euclidean space R^n with "n" dimensions, which is known as the "space of the results".[100]

When throwing a die, for example, the space is one-dimensional and the admissible results may be associated with the whole numbers from 1 to 6. Identifying a point of the earth's surface, on the other hand, calls for a two-dimensional space (latitude and longitude).

An "event" corresponds to a set of points in the space of the results. For example the event "result > 3" in the throw of a die corresponds to the set {4,5,6}, the event "ROME" corresponds to the set of points on the territory belonging to the city of Rome.[101]

The space of the results is considered as being outside the scope of the survey and is therefore not represented.[102]

A generic modality, insofar as it corresponds in probabilistic terms to an event, can be considered a "set" precisely in reference to the points in the space of the results. The extension of this set does not fall within the field of observation, given that the space of the results is not represented. (see §2.7).

[100] Also known as the "sample space", see [35] p. 125.
[101] If time is also considered, the space of the results on which the "statistical event" is defined is the combined space relative to the random variable concerned and the time of reference.
[102] This option is orientated to the economy of representation.

2.14 The space of events and its representation

Events, like results, can be associated to points in a Euclidean space of "m" dimensions, which is called "space of the events". Each point of the space of the events represents an event, i.e. a set of points of the space of the results or, in other words, an element of the power set of the space of the results.[103]

The choice of dimension "m" is largely arbitrary and guided by the economy of the representation. In practice, each aspect of reality that we intend to consider separately (e.g. territory, category of economic activity, etc.) is linked to a single dimension and multidimensional spaces are obtained by "joining together" one-dimensional spaces (see §2.17). For example, for the identification of a place, the space of the results is of dimension "2" (see §2.13), while the space of the events can be assumed to be of dimension "1" (provided there is no need to distinguish between longitude and latitude).

It must also be granted that in different situations the perception of reality and, therefore, the choice of dimension may be different, and that spaces with different structures but of equal significance may co-exist even within the same field of observation (for example, in the above case, longitude and latitude could also have been represented separately in two discrete spaces of events, in which case the space of the places would be two-dimensional).

The algebraic structure of the space of the events (Boolean algebra) is often referred to in the literature as the "algebra of events".[104]

The "certain" event represents the set of all possible results of a trial (the entire sample space) and corresponds to the universe of the statistic set. All other events correspond to subsets of the universe.

Atomic (or elementary) events are those that contain only one possible result. They cannot be broken down in terms of other events.

Each event of the space can be expressed as the union of specific atomic events. For example, the event "youth" can be defined as the union of the ages between 0 and 17. The union of all the atomic events is the certain event.

Unlike the space of the results, the space of the events is considered internal to the field of observation and is represented as a "power set". The intension defines the types of admissible events, the extension consists in listing the elements corresponding to the events "of interest".[105]

Each event can be defined by means of a statement that refers to the space of the results. For example, for the chance phenomenon consisting in determining the age of people, if the space of the results consists of the continuous space of the "ages", the events "young", "mature", old" and the years from 0 to 100 can be defined in the space of the events in terms of intervals of instants of age.

[103] The space of the events therefore corresponds to a power set.
[104] See, for example [35] pages 123 ff.
[105] Apart from being burdensome, it would in some cases be impossible to represent all the admissible events of the space, because they are infinite in number (for example, in the case of spaces relative to continuous random variables); the cardinality of the extension, for obvious practical reasons, is instead always finite.

The "certain" event is the only event whose explicit representation is considered essential;[106] it corresponds to the entire space considered unitarily, in other words not disaggregated.

The possible representation of atomic events, notwithstanding that the space of the events is external to the field of observation, makes it also possible to consider single results ("atomic events" are sets of a single element, consistently with the statistical context, see §1.2). However, it may not be possible to represent all the atomic events explicitly,[107] or it may not be of interest to do so.[108]

Set relationships are established between the events explicitly represented (in other words "existing"). Of particular interest is the representation of the relationship of aggregation/disaggregation, for the reasons already mentioned (see §2.9).

Existing events can be subdivided into "compound" and "simple", according to whether or not they can be expressed in terms of the union of other existing events. "Atomic events", if represented, are simple.

The possible non-representation of atomic events leads to the loss of useful information. The expression of each event as the union of atomic events makes it possible to determine whether or not two generic events are disjointed.[109]

This information can be retrieved when the simple events are all disjointed, when they correspond to a partition of the certain event (to the greatest detail that is of interest) and when every other event of interest is expressed (either directly or recursively) as the union of a sub-set of those events. In this case, it is possible, on the basis of the set relations between events, to determine whether or not two generic events are disjointed.[110] There may also exist practical cases in which it is of no interest or it is inopportune to satisfy the conditions set out.

Simple events correspond to the greatest level of detail that it is intended to represent and they may change in relation to the observer's interest. If, for example, greater detail on the territory (the city, rather than the nation) becomes desirable from a certain reference time on in a periodic survey, it could be necessary to define a new set of simple events (cities, presuming that the simple events considered previously were nations).

The process of formalizing the space of events corresponds to an operation commonly known as "coding". This consists in linking each "existing" event to a symbolic code that identifies it (the spaces of events considered in the theory are numerical, but for representational purposes, it is equivalent of using alphanumerical spaces of events).

Given a set "C" of symbols and a set "E" of events, we will thus speak of a "coding system" or "coding function" as a function, $f_c: C \rightarrow E$, which links each code "c" of the set "C" to a single event "e" of the set "E". The set "C" (the "domain" of "f_c") is a "statistical set" and its "elements" are, precisely, the "codes". The set "E" (the co-domain) is composed of the statements that define the statistical events.

---

[106] The existence of only the certain event in the space of the events is generally an extreme case that is of no use.
[107] For example, for continuous variables they are infinite.
[108] The information at a lower level of detail may be considered sufficient.
[109] Two events that can never occur simultaneously are disjointed; this happens when their intersection is the empty set and they thus have no atomic events in common. There are no redundancies of information between statistical data relating to disjointed "elements".
[110] If the relations are represented formally, this deduction can also be automatic.

In case the same set of events is codified more than once (for example, by different coders), it may be of interest to represent suitable relationships of synonymy to link the elements with the same meaning in the different coding systems.

2.15 Sub-sets of the space of events

Many random variables may refer to the same "results space" and the same "events space".

One of the aims of administering statistical concepts is to bring back phenomena relating to the same aspect of reality (e.g. the territory) to the same space of events, given that it is the basis for comparing the related data.

As an example, the variables "city of birth", "nation of residence", "continent of birth", "nation of export", and so forth may refer to the same space of events "areas of territory".

The admissible events of a random variable are usually a sub-set of events "existing" in space.

In the most intuitive case, they belong to a "partition" set, but cases may vary greatly (e.g. a union of "partition" sets and "isolated" events", sub-spaces of events, also coinciding with the whole space …..).

For example, the admissible events of the variable "nation of residence" are the elements of the set of the nations, which is a partition of the territory. The admissible events of the variables "age-class", "years of age" (see also §2.14), i.e. respectively {"young", "mature", "old"} and the set of years, also correspond to partitions.

In the case of the two variables in a classic double-entry table, on the other hand, the admissible events are the union of a partition and its universe (for example, the single nations and the whole world).

More generally speaking, therefore, the admissible events of a variable may be represented by a generic set of events, which is a sub-set of a space of events.

Sub-sets exist in relation to time and may contain any event that exists in the space of events. An event may also be part of a given sub-set for a period of time included in that of its existence.

2.16 Partition and classification sub-sets

Among the sub-sets of a space of events, particular importance attaches to those that correspond to partition sets, whose elements therefore form a "set disaggregation" of a specific event (the universe of the sub-set).[111]

An event can be disaggregated in several ways, generating numerous set disaggregations at differing levels of detail. There may even be no significant relations between them.[112] Often, however, it is of interest to consider the levels of detail in hierarchical order. For example, when treating the territory, the levels relating to continents, nations and cities: each continent is an aggregation of nations, each nation an aggregation of cities.

When effecting a hierarchical breakdown, because the partition operator is applied[113] to both the events and the set of events, various types of relations are established between sets of events, sets, and between events and sets.

An initial type of relation is that which links events belonging to two consecutive levels of detail. This is the set aggregation/disaggregation relationship[114] (e.g. Europe, a "continent"-level event, can be broken down into Italy, France, …., which are "nation"-level events).

A second type of relationship links an event to the set of events of which it is the universe (e.g. Europe is the universe of the set "Nations of Europe", which contains the events Italy, France, ……).

A third type of relationship links a set of events to a group of its disjointed sub-sets of which it is the union. This is the "classification" relationship, which links a set of events to one of its partitions (e.g. the set of "nations of the world" can be broken down into the sets of "nations of Europe", "nations of Africa", ….).

It should be noted that it is also possible to classify sets that contain non disjointed events. If, however, the events of the set that is classified are all disjointed (as in the above example), the universes of the sets will correspond to a disaggregation of the universe of the original set.

[111] It may coincide or be included in the universe of the space of events.
[112] For example, the Italian territory can be disaggregated into cities, postal districts, religious dioceses, between which there may be no relationship of aggregation/disaggregation.
[113] Breakdown of a set into disjointed sub-sets whose union supplies the initial set.
[114] See also §2.9.

2.17 Spaces of combined events

Spaces of events of more than one dimension (in which multidimensional random variables assume values)[115] may be defined by joining together one-dimensional spaces.

Given the chance phenomena "F_a" and "F_b", the combined phenomenon "F_c" consists in the combined execution of F_a and F_b. A generic result of the combined phenomenon is formed by an ordered couple of results of F_a and F_b, i.e. $r_c = (r_a, r_b)$. If R_a and R_c are the respective spaces of the results, the space of the results of the combined phenomenon is $R_c = R_a \times R_b$. A generic combined event is therefore a sub-set of Rc = Ra x Rb and the space of the combined events Ec is the power set of Rc, i.e. Ec = P(Rc). The combination of more than two chance phenomena can be defined in the same way.

With regard to the combined phenomenon, the phenomena Fa and Fb are known as marginal. Similarly, a generic event of the combined phenomenon that occurs only when a given event of "Fa" (or of "Fb") occurs is known as marginal.

It should be recalled that the domain of the statistical function is generally multidimensional (see §1.3) and thus belongs to a combined space.

Before any combined space can be defined and represented it is necessary to define and represent the one-dimensional marginal spaces of which it is composed. Any space obtainable by combination of the existing one-dimensional spaces can be assumed as implicitly defined as above.

In the space of combined events those of interest (i.e. existing) by default are all the events belonging to the Cartesian product of the events existing in the one-dimensional spaces of which it is composed.

The preceding assumptions, which should simplify the definition process, may however, not be sufficient to identify every combined event that is of interest.

The Cartesian product of two spaces of marginal events is, in fact, included in the space of the events of the combined phenomenon.[116]

The events of the combined space can thus be divided into two types: "Cartesian" (i.e. corresponding to elements of the Cartesian product of marginal events) and otherwise.

An example of a Cartesian event, in the combined space formed of the "zones of territory" and "types of economic activity" is the event:

- (zone = Europe and activity = industry).

[115] Vectors of one-dimensional random variables.

[116] If E_a, E_b are the marginal spaces and E_c the combined space, if R_a, R_b, R_c are the respective spaces of the results, then $R_c = R_a \times R_b$ and $E_c = P(R_c) \supset E_a \times E_b = P(R_a) \times P(R_b)$.

An example of a non Cartesian event is the following:

- ((zone = Western Europe and activity = engineering industry) or (zone = Eastern Europe and activity = textile industry)).

Non-Cartesian events can generally be expressed as the composition of Cartesian events (as in the above example).

However, in order to preserve the same multidimensional representation structure of the Cartesian events, it is preferred to proceed so that non-Cartesian events also correspond to elements of the Cartesian product of the one-dimensional marginal spaces. To this end suitable events conditioned[117] by events of other marginal spaces are defined in one (or more) of the marginal spaces.

To return to the above example, if in the space of events relating to the territorial zone we define the events:

- X: (zone = western Europe $\mid_{\text{activity = engineering industry}}$)
- Y: (zone = eastern Europe $\mid_{\text{activity = textile industry}}$)
- Z: $(X \cup Y)$

the non-Cartesian event in the multidimensional space can be expressed as:

- (zone = Z and activity = industry).

[117] Event A is conditioned by event B if A is considered to have occurred only if B also occurs. This conditioning is indicated by the symbol $A \mid_B$.

2.18 Sub-sets of spaces of combined events

The admissible events of a multidimensional random variable can be represented as a sub-set of the Cartesian product of the spaces of the events of the one-dimensional component variables (see §2.17).

The representation of the extension of a generic sub-set of a combined space would require a detailed list of the elements of the Cartesian product of which it is composed.[118] In the case in question, however, it is usually both possible and convenient to use a more summary representation.

The extension of the sub-set of the combined space often coincides with the Cartesian product of suitable sub-sets of the marginal spaces (the domains of definition of the one-dimensional variables).

Where this is not the case it is generally preferable to define a Cartesian product of marginal sub-sets that includes all the events of the sub-set of the combined space. The combinations to be excluded (or the complements to be included) can be identified by means of additional Cartesian products,[119] defined in the combined space or even in its sub-spaces.

It is, in particular, possible to define the combinations of marginal events that cannot occur (incompatible events), using conditions expressed in reference to the only marginal spaces involved. In the same way it is also possible to define, if necessary, the compatible combinations.

The conditions of incompatibility that express the interdependence of events belonging to specific spaces are inherited by each vector of variables that refer to the relative combined space. Conditions of incompatibility may also exist only for specific vectors of variables.[120]

The most obvious examples of compatibility or incompatibility are those relating to the existence and admissibility of events in relation to time. Existence induces conditions of compatibility between the time and the events of another generic space. The admissibility (of an event for a variable) induces conditions of compatibility between the time and the admissible events of a generic variable.

[118] The sub-set of a Cartesian product is known as a “relation”.
[119] Which can also coincide with single combinations of events.
[120] The following section mention will also be made of incompatibilities that emerge only within the ambit of specific statistical functions (see §3.8).

2.19 Conclusions

In closing this chapter, the following is a brief summary of the general approach to the representation of concepts.

The basis consists of the representation of the spaces of one-dimensional events corresponding to the fundamental categories of reality (domains). The categories exist in relation to the requirements of the specific field of the survey (for example territory, economic activities, subjects, duration and so on). The category of the "reference time" is expected to be necessary,

The general structure of a space of events is the Boolean algebra, which makes it possible to represent, in the same space, events corresponding to different levels of detail, as the statistical context requires. Each space of events gathers the definitions of the events of interest and of the set relations that are established between them, which trace the explicit structure of the space.

The other sets of events that are of some interest (such as those in which the variables assume values) are defined as sub-sets of the spaces of events.

The spaces of multidimensional events obtained by combining one-dimensional spaces are considered as implicitly existing. The sub-sets that are of interest are explicitly represented and express relations between marginal spaces (e.g. domains of definition of combined variables, incompatibility of events, …).

To each space of events is associated a certain number of variables, which qualify the different meanings with which the events of that space can be interpreted. Each variable has a set of admissible events.

The basic types of concept described by the dictionary of concepts are thus the "events" (or "elements"), the "sets of events" (spaces of events and their sub-sets) and the "variables", as well as their reciprocal relationships. The premise for a formal representation is that each concept be codified, in other words identified by means of a symbolic code.

Suitable attributes (meta-information) are used to describe types of properties relating to the different types of concepts and relationships.

3. THE INTENSIONAL FORM OF STATISTICAL DATA

3.1 Generalities

The "statistical function" introduced in the first chapter[121] can be considered the algebraic equivalent of the notion of "statistical datum" typical of descriptive statistics.

A generic function can be described in two forms: intensional and extensional. The one corresponds to the definition of the function, the other is the extended expression of the associations between the elements of the domain and those of the co-domain.

The aim of generic statistical investigation is knowledge of a complex of functions. The intensional definition serves to specify "a priori" the objective of the investigation, whereas the extensional form constitutes the final result.

The focus of this chapter is on the intensional form of statistical functions[122] and the correlations between this form and the extensional form.

As regards the statistical data, moreover, knowledge of the procedure by means of which they are obtained is found to be useful and important. Basically, there are two ways:[123]

- the taking of measurements;[124]
- the processing of other statistical data.

The last part of the chapter addresses the question of the formal representation of the algorithms (rules) for the calculation of data starting from other data, which appears vital for the purpose of automating the processes within the information system.[125]

[121] A law that associates a mode of a statistical characteristic to every pair consisting of a group of statistical units and a mode of time (see §1.3).

[122] The "dictionary of the statistical data" serves to record the intensional form, which essentially describes the static structure of the datum (see §1.6).

[123] See [45] p.19.

[124] Which, in turn, can be direct, where they are taken on the system under investigation, or indirect, where they are obtained from other information systems or other statistical observations.

[125] The formal definition of the rules of calculation within the dictionary of data permits their "active" use by the software.

3.2 The statistical function

Given two sets A and B, a "statistical function" is a law that associates one and only one element $b \in B$ with each element $a \in A$. The set "A" is called the domain of the function, the set "B", the co-domain. A single association is normally known as an "occurrence".

The sets A and B belong to spaces identified by multidimensional variables, obtained by the conjunction of appropriate one-dimensional variables. The statistical function is thus characterized by a vector of independent variables[126] that assumes values in the domain and by a vector of dependent variables[127] that assumes values in the co-domain.

Since we are dealing with a statistical context, it needs to be recognized that the variables, the sets and their respective elements are statistical concepts,[128] and that in turn the elements represent sets of real-world objects.[129] The variables, moreover, can be either quantitative or qualitative.[130]

The combinations of values of the independent variables (the elements of the domain) identify the groups of objects that are to be described (e.g. groups of persons resident in the same country) and the time to which the description refers. The dependent variables correspond, instead, to the types of properties described (e.g. income). Nothing prevents us from considering groups made up of single statistical units and types of properties related to them.[131]

Lastly, from the point of view of the information it expresses, every occurrence of the function can be considered an elementary "message"[132] corresponding to an affirmation of the type: "The group $\mathbf{g_i}$, a subset of the set of statistical units **G**, with reference to time **t** is characterized by the manifestation $\mathbf{c_i}$ of the characteristic **C"**.

[126] Corresponding to the systematic characteristics and to the reference time.
[127] Corresponding to the statistical characteristics.
[128] In particular, they are historical and correspond to a series of algebraic objects, each relative to a single instant (see §2.4).
[129] See §1.2 and 2.7.
[130] See §1.1, [33] pp. 11-15 and [35] pp. 28-29.
[131] See also the difference between the variables of the "object characteristics" and those of the "statistical characteristics" in [47] p. 23 and [50] pp. 2-4.
[132] See [45] pp. 5-7, [46] pp. 15-24, [47] pp. 7-11, and [50] pp. 2-4.

3.3 The probabilistic approach

According to the probabilistic approach discussed in the previous chapter (see §2.12 ff), the statistical function can be considered the result of a series of tests relative to a particular chance phenomenon.

Given an urn composed of elements of the function's domain (pairs consisting of a group of statistical units and a manifestation of time), consider first the random event consisting in the extraction of an element from the urn and the determination of the modes that occur in relation to both the "vector" of dependent variables and the vector of independent variables. This is a combined chance phenomenon, compared with which the two chance phenomena consisting in the separate consideration of the dependent variables alone and the independent variables alone are known as "marginal".[133]

Consider now the conditioned chance phenomenon in which the conditioned variables are the independent ones. The conditioning event, that is, consists in the extraction of a particular group-time pair, while the conditioned event is the vector of modes that occur for the statistical characteristics.

The extensional form of the statistical function can be obtained by conducting a series of consecutive tests on the conditioned event, without putting the extracted element into the urn, until the urn is empty.

It should be noted that filling the urn again and conducting another series of tests makes it possible to obtain additional extensional forms, characterized by the same conditioning events, but in which the conditioned events may change owing to the random factors involved. In other words, with every reiteration of the series of extractions we obtain another statistical function, characterized by an intensional form equal to that of the earlier tests and by an extensional form that may be different.[134]

[133] See §2.17.

[134] The technique of data "redundancy" is adopted in the practice of statistical observations; with the aim, for example, to controlling observation errors.

3.4 Groups of statistical units and time

In the most general case, a statistical function refers both to many groups of statistical units and to many manifestations of time.[135]

Certain types of statistical data are nonetheless distinguished in relation to the groups and times considered:

- single occurrences, i.e. data whose domain is relative to a single manifestation of time and a single group of statistical units;
- cross-section data, whose domain is relative to a single manifestation of time and many groups of statistical units;
- historical series, whose domain is relative to a single group of statistical units and many manifestations of time.

In a cross-section, the extensional form is a set of occurrences referred to the "same" time, while in a historical series it is a set of occurrences referred to the same "group of statistical units".

In the most general case, the extensional form of a statistical function can be considered alternatively as:

- a set of occurrences (each of which relative to one group of statistical units and to one manifestation of time);
- a set of cross-section data (each of which relative to one manifestation of time);
- a set of historical series (each of which relative to one group of statistical units).[136]

In the most general case, furthermore, the groups of statistical units of the function domain may vary with time. So, in abstract terms, the cross sections of the same function may also be relative to different groups and the historical series of the same function may also be relative to different manifestations of time.

[135] See also §1.1.
[136] See also [38].

3.5 The intensional definition

The intensional form is at the same time the definition of the existence and significance of the function, as well as the algebraic definition of the latter's structure.

It can be considered as made up of various sub-definitions (described in more detail in the following sections) which comprise:

- the identification of statistical function by means of a symbolic code and the verbal description of its significance;
- the specification of the dependent and independent variables that define the space within which the function can be represented;
- the specification of the domain and the co-domain (sets of the admissible combinations of the values of the variables).

It has to be recognized that every intensional form corresponds to a specific extensional form, which, however, may be partly or wholly unknown, and hence not represented. In fact, the function can be defined even before the extension is known (see also §1.6). Conversely, it cannot exist an extensional form in the absence of the corresponding intensional form.

The intensional form refers to statistical concepts.[137] So, there are relationships between functions and concepts which induce specific integrity constraints. These have to be verified in relation to every instant of time and therefore require, for it to be possible for a function to refer to a concept, that the latter is considered as "existing" in the field of observation[138] in relation to the time at which it is referenced.

The intensional forms make up the schema[139] of the statistical data and provide a documentary basis that can be used by different people and for different purposes. For example, by the designer of a statistical survey as the "specification" of the desired data, by the producer of data as the "specification" of the data to be produced, by the administrator as a control instrument, by the user as documentation of the data contained in the information system, and so on. In "active" systems it is also used by the software.[140]

[137] In particular, to the variables and related sets of admissible values.
[138] I.e. that it is defined in the dictionary of concepts.
[139] The result of applying the representation model to the specific case.
[140] For example, by formal control functions to verify the correspondence between intensional and extensional forms, by calculation functions to determine the transformation algorithm between inputs and outputs, by data search and presentation functions, and so on.

3.6 The one-dimensional variables of the function

In general a statistical function is characterized by "n" independent variables and "m" dependent variables,[141] which configure the space in which the function is represented.

A function with "m" dependent variables can be turned into the case of a vector of "m" functions, each of which is relative to just one of the variables. It is thus possible without loss of generality, to treat only the case of functions with a single one-dimensional dependent variable.[142]

The role the (dependent or independent) variable plays is a property of the association between variables and functions. In fact, a variable may play different roles in different functions.[143]

Ideally, independent variables can be divided into two categories. The first identifies the groups of statistical units, the second the manifestations of time.

The category relative to time contains a variable[144] that specifies the "reference time".[145]

The category that identifies the "groups", that are the subsets of the statistical class to which the function refers, contains one or more variables.[146]

There can be several different methods to identify the groups, for example:

- using a single variable, relating to the statistical class of the datum, whose values correspond to the groups of the grouping;
- using one or more variables, the combinations of the values of which correspond to the groups and represent the properties that the statistical units must possess (the criteria for belonging to the groups);[147]
- using a vector of variables, the combinations of the values of which constitute a unique identifier of the groups (the "name" of the group)[148].

In a sense such methods are equivalent since the name of the group the statistical units of the class belong to can be considered one of their properties.

---

[141] Understood to be one-dimensional.

[142] To which corresponds a single statistical characteristic.

[143] For example, the variable "form of government" is dependent in the function "form of government of nations" and independent in the function "income of nations by form of government".

[144] In reality, there can be more than one time variable (see [47] pp. 8-11 and [50] p. 4). For the sake of simplicity, the analysis is restricted to the case of just one (see also §1.3), which is the most common one.

[145] The reference time is considered to exist even in the case of cross-section data and to be characterized by a single admissible value which is the same for all the observations.

[146] Identifying the groups of statistical units also entails the identification of the statistical class.

[147] Each group, which is a subset of the statistical class (e.g. persons), contains the statistical units that have the combination of values characteristic of the group (e.g. age = 20 years and gender = male and country of residence = Italy); see also §1.3.

[148] The first and the third methods are quite similar: in the former the "name" of the group is one-dimensional, in the latter it is multidimensional; such methods can be used, for example, when consideration is given to groups of predetermined statistical units (see also §1.1 and 1.3) the composition of which is specified by listing the single statistical units and, in particular, in the case of groups made up of single statistical units, which can be identified by the name of the statistical unit of which they are made up.

As for the dependent variable (which corresponds to the statistical characteristic), two types can be distinguished:[149]

- the variables that express properties relative to groups of statistical units of any cardinality (e.g. "total income" or "average income");
- the variables that express properties relative to groups made up of single statistical units (e.g. age, country of residence, …).

The latter case is characteristic. in particular, of data referring to objects originally conceived as "elementary" and that become part of the "statistical" context (see §1.2).

[149] See also the difference between the variables of the "objective characteristics" and those of the "statistical characteristics" in [47] p. 23 and [50] pp. 2-4.

3.7 The domain of the one-dimensional variables

Each variable in the field of observation is characterized by the set of values that it can have (the "definition domain" of the variable).[150] When it is used in connection with a specific statistical function, however, it is recognized that the admissible values may also be a subset of its definition domain.

The set of admissible variables for a variable in connection with a specific function is called "domain in use" and may coincide with the definition domain or be a subset thereof. For example, the definition domain of the variable "country of residence" consists of all the countries of the world, but its domain in use in connection with the function "income of Europeans by country of residence" is made up only of the European countries.

The domain in use is a property of the association between functions and variables. It is defined by referring to a subset of the definition set of the variable, which, in turn, is defined in the dictionary of concepts.[151]

In special cases, a variable's domain in use consists of a single element. In this case we speak of "non-disaggregated" variables.[152]

An example of this is the domain in use of the "reference time" in the case of cross-section data. In the case of historical series, by contrast, the domain in use of the "reference time" is the only one to contain more than one value.

In general, a domain in use can contain events with different levels of detail.[153]

This occurs, for example, for the variable "area of residence" of the function "average income of persons", where they involve cities for persons resident in Italy and countries for persons resident abroad.

Moreover, events that are not disjointed may belong to the same domain in use. With reference to the previous example, this occurs if the areas considered are both countries and continents.[154]

Another example can be found in so-called double entry tables (i.e. tables having two systematic characteristics) that also include row and column totals and the overall total. In such tables four levels of detail are present at the same time, obtained by grouping the statistical units in the following ways:

- with respect to the joint change in two characteristics;
- with respect to the change in the row characteristic alone;
- with respect to the change in the line characteristic alone;
- in a single group coinciding with the entire statistical class.

[150] See §2.8.
[151] A given "domain in use" may be referred to more than once, by different variables or different functions.
[152] If the variable is the dependent variable, the function is constant.
[153] Marginal events with different levels of detail are useful in identifying groups of statistical units with different levels of detail.
[154] Marginal events that are not disjointed are useful in identifying groups of statistical units that are not disjointed.

The generic statistical unit belongs simultaneously to four sets of the grouping, which, therefore, are not generally disjointed.

The admissible events of each marginal variable belong to two levels of detail (a partition set and its universe), the combination of which generate the four levels of detail referred to above in the combined space.

3.8 The definition domain of the function

The definition domain of the statistical function is a subset of the Cartesian product of the domains in use of the independent variables and thus belongs to the combined space of events identified by these variables.

The definition domain often coincides with this Cartesian product. In such cases, knowing the domains in use is sufficient to determine the domain of the function.

In the opposite case, the combinations of values that are not part of the domain can be specified by means of conditions of mutual incompatibility (or compatibility) between the values of some (or all) of the variables, which can be expressed as Cartesian products[155] in the subspaces identified by the variables in question.[156]

The clearest example of compatibility is that concerning the admissibility of events in the domains in use of the variables of the function.[157] It produces compatibility constraints between the reference time and the admissible events of other variables of the function, which condition the definition domain of the function itself.

Other compatibility or incompatibility constraints include those that can be established between events that belong to different spaces[158] or are admissible for different "non-temporal" variables (see §2.18).[159] In particular, the domain of a function inherits the constraints, if any, that act on the spaces of events or on the variables to which the function refers.

There can also be incompatibility constraints relative to the specific statistical function, that depend on the information that the function must provide rather than the abstract schema of the system under observation.

For example, there may be constraints deriving from the choice of statistical class. A hypothetical statistical class made up of adults and employed minors, for example, would produce a mutual incompatibility between the variables (if present) "age class" (young, middle-aged, old) and "employment status" (employed, unemployed), because the combination of "young" and "unemployed" would prove not to be admissible in the domain of the function.

The choice of the groups to be considered can also produce incompatibilities. For example, the choice of groups made up of residents (in Italy) born in the same city and non-residents born in the same country would produce a mutual incompatibility, in the ambit of the domain of the function, between "residents" and "countries" and between "non-residents" and "cities".

[155] In the extreme case, coinciding with single combinations of events.
[156] See also §2.18.
[157] A particular event may be admissible in a domain in use for a time that is shorter than those of its existence or its admissibility in the definition domain of the variable.
[158] These constraints depend on how the events are actually defined.
[159] These constraints depend on the specific definition of the variables.

3.9 The search of a common space

Independent variables whose domain in use consists of a single value (see §3.7) are known as "non-disaggregated" variables. In addition, variables that are represented in extensional form are known as "explicit" and the others as "implicit".[160]

The two notions are not independent; in fact, disaggregated variables must necessarily be explicit. On the other hand, while non-disaggregated variables are generally implicit, they can also be represented explicitly.

The intensional form of a statistical function must undoubtedly contain the definition of all the "explicit" variables. It is also recognized that it may include the definition of implicit, i.e. non-disaggregated, variables. The latter's usefulness derives from the need to bring different functions back to the same multidimensional space in order to permit their comparison and joint utilization.

Suppose, for example, that one wishes to compare "personal income by country of residence and marital status" (f_1) with the "income of residents in Europe by marital status and sex" (f_2). In this case it is necessary to bring the two functions back to the same combined space including the variables of both f_1 and f_2 (i.e. "country of residence", "marital status" and "sex"). With reference to this space, in the case of f_1, "sex" (whose domain in use consists of the certain event, i.e. the union of the two sexes) is implicit, while in the case of f_2, "country of residence" (whose domain in use consists of the "Europe" event, i.e. the union of the European countries) is implicit.

Lack of knowledge of the implicit variables would preclude, at least as regards formal representation, the comparison or joint use of the two functions.

For the purpose of formalized construction of a common space, it is possible to avoid the representation in the intensional form of implicit variables whose admissible event corresponds to the universe of the definition domain of the variable. The aim of this is to reduce the number of definitions, which, where there are many functions (to be rendered mutually comparable), can become extremely burdensome. The addition of a new function with a new variable, for example, would imply going back over the definitions of all the existing functions that were to be rendered comparable with the new one.

[160] Typical examples of implicit variables are "time" for a cross-section datum and the variables that identify the "group" for a historical series. It should also be noted that cases of implicit variables can arise whenever a variable disappears in the extensional form as a consequence of the aggregation of all its values.

3.10 Attributes of the statistical datum

The definition of the intensional form can be supplemented by adding a series of descriptive attributes[161] that represent metainformation on functions or on the associations between functions and variables.

The choice of a satisfactory set of attributes is part of the activity of designing each data dictionary. The balance between the operating costs of a particular piece of information and the benefits it can bring is case-specific and depends also on the level of automation available. Typical examples of attributes include the distinction between stocks and flows, physical size (e.g. cardinalities, amounts, rates, etc.), the unit of measurement (e.g. the Italian lira), the scale (e.g. millions), periodicity (e.g. quarterly data) and so on.

In addition to the properties described by its own attributes, the generic function also inherits the characteristic properties of the variables, sets and elements that contribute to its definition. Such properties are described by attributes in the field of observation (see §2.8).

On the other hand, certain attributes can vary as a function of the single occurrence of the extensional form. In such cases, the corresponding attribute must be defined in the intensional form in the same way as the other variables[162] and represented explicitly.[163] For example, suppose that the "currency" in which "personal income" is expressed can vary according to the individual occurrences. It becomes essential, if the information is to be understood correctly, to qualify each occurrence with information on the currency in which income is expressed.

Properties can also exist that vary as a function of subsets of the definition domain of a function.[164] In such cases, it may prove useful to associate the corresponding attributes with the definition of the subsets, which, as mentioned earlier, can be done by referring to suitable marginal subspaces (see §2.18 and 3.8).

[161] An attribute is a type of property.

[162] This constitutes another type of variable, the descriptive variable (see also "attributes" in [38]).

[163] A similar metainformation can also be considered as the dependent variable of a suitable function, whose domain consists of the occurrences of the statistical function and which, for the sake of convenience, is represented jointly with the latter's extensional form.

[164] One example is given by properties dependent on the single historical series or the single cross-sections of which a function, in its general form, can be considered to consist (see §3.4).

3.11 The knowledge domain of the function

The knowledge domain of a function is the subset of the definition domain for which the extensional form is available (i.e. in which the function is "known"). It is "empty" for functions for which the extensional form is not available.

The knowledge domain can be specified using the techniques used to specify the definition domain (see §2.18 and 3.8).

Knowledge of the function may also depend on single variables.

A common case is time. For example, in periodic statistical observations the functions, even though they are defined over a longer interval (the period in which the observations are to be made), are generally known with reference to the times at which the observations have been made.

Another common case is the collection of data from a large number of "reporters". In this case knowledge of the function depends on the manifestations of the "reporter" variable (some reporters may have already made their reports, while others still have to make theirs).

The two preceding criteria can be simultaneously valid, and in this case the knowledge domain is determined with reference to the combined space.

In reality, the knowledge domain could also be deduced from the "extensional form" in the case in which provision was made for the complete development of the combinations of values of the variables for which the function is known. The absence of a combination of values would imply that for that combination the function was unknown.

However, the separate representation of the knowledge domain, apart from being more synthetic and comprehensible, does not impose constraints on the choice of the manner of representing the extensional form, which can therefore be adapted to the needs of individual cases.

In fact, the extensional representation of the knowledge domain is not always practical or economically advantageous. Even where there are only a few independent variables, the cardinality of the domain may be so high as to create operational difficulties (it is sufficient to have 10 classification variables with a domain of cardinality 10 to develop 10 billion combinations).

In such cases, however, for the majority of the elements of the domain, the function is often found to have a constant value that is typically equal to zero (i.e. it can be likened to a sparse matrix that has the same value almost everywhere in the domain) and it may be better not to represent all the combinations. If the decision is taken not to represent the function explicitly when it is equal to zero, the knowledge domain would no longer be deducible from the extensional form (in the absence of a certain combination of values it would not be possible to deduce whether the function is unknown or equal to zero in correspondence with that combination).

Such representation choices depend not only on the cardinality of the domain but also on the efficiency with which the particular instrument used can memorize the sparse matrices (major progress has been made recently in this field with the introduction of so-called multidimensional DBMS).[165]

[165] See also [28].

3.12 The shift to the extensional form

The extensional form of a statistical function can be represented intuitively as an ideal "table".

Every row describes an occurrence of the function, i.e. the association between an element of the domain and one of the co-domain. Every column, by contrast, corresponds to a variable represented explicitly, of which three main types can be distinguished:[166]

1) the independent (or "key") variables of the function;
   - variables that identify the groups:

     the variables related to the statistical class which identifies the groups of statistical units directly;

     the variables corresponding to the grouping criteria, which give the properties that the statistical units must possess to belong to the groups;
   - variables that identify the times:

     the reference time;
2) the dependent variables (one or more depending on the structure of the function's co-domain);
3) the descriptive attributes (which represent the types of properties characteristic of the occurrences of the function).

Another way of representing the extensional form, which is more consistent with the mathematical concept of function, consists in imagining a space whose dimensions consist of the independent variables. Each point of this space is associated with a single value of each dependent variable and each descriptive attribute.

These representations correspond very well to two of the most widespread logical structures for representing data: i.e. relational tables[167] and multidimensional data bases. Both are suitable for containing statistical data.

The problem of defining effective structures for extensional representation nonetheless deserves to be analyzed much more extensively and is beyond the scope of this paper.

[166] For the sake of simplicity, no reference is made to possible forms of denormalization of the function, such as, for example, the introduction of variables functionally dependent only on a subset of the "key" variables.

[167] It should be noted that a "function" is a special case of the "relationship" considered in the "relational" model (see [13]) and that a "tuple" of this model is equivalent to a row of the above-mentioned table or a point of the above-mentioned space (i.e. to an occurrence of the function); nonetheless, whereas in an analytical context a "tuple" refers to an object conceived as "single", in a statistical context it identifies a group.

3.13 The calculation of statistical functions

The statistical data that are available can be used to calculate other data.

Calculations are of considerable importance for statistical information systems because they make it possible to obtain "semi-finished products" and "finished products", starting from raw materials: the "elementary" data the information system acquires externally.

The formal representation of the calculation algorithms (rules) is thus essential both for documentation purposes and for automating production processes.[168]

The process of formalization can be based on the concept of "transformation", i.e. a process that produces a "result" statistical function by applying an "algorithm" to one or more "operands", which are also statistical functions.[169]

The subject of the transformation is both the intensional and the extensional form of the operands, to produce the intensional and extensional form of the result.

Single transformations can be generically represented as relationships between the result functions and the related operands (one attribute of the relationship is the algorithm applied). In the space of the functions, such relationships form a tree of transformations that traces the system's calculation processes.

The algorithm describes the transformation of both forms (intensional and extensional) and is generally composite, since it combine elementary operators in an expression.

One problem characteristic of all transformations concerns the determination of the "calculability" of the result or, in other words, the determination of the knowledge domain of the result in relation to the knowledge domain of the operands. In general, a given occurrence of the result function can be calculated (and hence, after the calculation, is known) where all the occurrences of the "operand" functions involved in its calculation are known.[170]

As can easily be imagined, there is a vast and variegated range of problems concerning the calculation of statistical functions. A systematic examination of these problems is beyond the scope of this paper.

Apart from mentioning the existence and importance of this subject, the following paragraphs will do no more than present three elementary types of transformations,[171] to show how algebra can help in their formulation.

---

[168] A primary objective is to memorize calculation rules in the form of metadata that can be consulted by users and interpreted by software or, where there is not feasible or economically advantageous, to express the rules with direct reference to the software routines that perform the processing.

[169] In information system calculations, statistical functions represent the unit of transformation or manipulation.

[170] The various types of transformations can be characterized by specific forms of conduct in relation to the calculability.

[171] A broader range of cases and a formal language for the definition of the transformations of statistical functions are specified in [5].

3.14 Aggregation

Aggregation is perhaps the most common transformation in a statistical context, because it makes it possible to obtain more summarized information.

With reference to the dependent variable of a function, aggregation consists in deducing the value shown for the generic group "g" of statistical units, starting from knowledge of the value in correspondence with other groups of which "g" is a set-type aggregation (see also §2.9).

For example, given the number of inhabitants of the individual European countries (Italy, France, …), transformation by aggregation consists in calculating the number of inhabitants of Europe.

Aggregation is not always possible. In general, whether it is or not depends on the nature of the dependent variable. For example, "total income" can be aggregated, while "average income" cannot.

A function can be aggregated if, however two disjointed elements in the domain are chosen, it is always possible to infer the value of the dependent variable in correspondence with their union.[172]

For this to be possible, denoting the domain of the function by "A", the co-domain by "B" and the function by "f:A→B", there must exist an internal composition law "#" with respect to the set "B" such that the function "f" shows the following property:

$$f(a_i \cup a_j) = f(a_i) \# f(a_j)$$

for each pair (a_i, a_j) such that $a_i \cap a_j$ = empty set.

Putting $a_k = a_i \cup a_j$, this property makes it possible to obtain the $f(a_k)$ from knowledge of the $f(a_i)$ and $f(a_j)$ by means of the composition law #. In fact,

$$f(a_k) = f(a_i \cup a_j) = f(a_i) \# f(a_j)$$

In practice the most common case is that in which the composition law is the sum, so that:

$$f(a_k) = f(a_i \cup a_j) = f(a_i) + f(a_j) \text{ with } a_i \cap a_j = \text{empty set}$$

It should be noted that a function can also be aggregated only in relation to some dimensions.

The most obvious case is the "time" dimension, with respect to which aggregation does not always make sense. It does not make sense, in fact, for stock data, which refer to single instants in time, while it does make sense for flow data, which refer to periods. For example, the number of

[172] Functions that can be aggregated are especially useful, precisely because they make it possible to observe and conserve information at the highest level of detail of interest and then to calculate it at any level of synthesis.

inhabitants of countries (stock data) cannot be aggregated with respect to the reference time, but can be aggregated with respect to the variables that identify the groups.

Moreover, the transformation consisting in aggregation can be specified and effected on each of the single dimensions with respect to which the function can be aggregated. In fact, two multidimensional events, a_i and a_j are disjointed if the respective marginal events are disjointed, dimension by dimension, and the marginal events of $a_k = a_i \cup a_j$ correspond, dimension by dimension, to the union of the marginal events of a_i and a_j.

The result of the aggregation is a new function characterized by the same variables as the operand function and by domains in use containing the aggregated events (for the variables with respect to which the aggregation is effected). Moreover, some explicit variables of the operand function may be transformed into implicit variables in the result function, where the relevant domain in use reduces to a single element.

The aggregation algorithm can be determined on the basis of knowledge of the composition law "#", of the intensional form of the function to be aggregated (derivable from the dictionary of data), of the intensional form of the "result" function (specified by the person who defines the aggregation to be made), of the relationships of set-type union between the events to be aggregated and the aggregated events (derivable from the explicit algebraic structure of the relevant event space defined in the field of observation).

3.15 Set-type operations

Set-type operations are transformations that act basically on the domain of the operand functions.

The subset of a function f_1 is a function f_2 defined in a subset of the definition domain of f_2, and coincident with f_1 in that subset.

The algorithm therefore consists simply in the specification of the definition domain of f_2 (subject to the constraint that it be a subset of that of f_1).

Particularly significant cases of subsets include, for example, the extraction of a "historical series" or a "cross-section" from the most general form of a statistical function.

The classification of a function "f" is a set of functions "f_i", with i = 1, 2, …., whose definition domains form a classification of the domain of "f" and such that each of them coincides with "f" in its own definition domain.

For example, the decomposition of a statistical function into historical series or cross-sections is a classification.

The union of functions can be considered the inverse of classification.

The merger of historical series relative to various groups of statistical units and the merger of cross-sections relative to different times are cases of union.

Union, however, is possible only where certain constraints are complied with. The operand functions, for example, must coincide in correspondence with the non-empty intersection, if any, of their domains. Furthermore, where the operand functions refer to different levels of detail, the latter must be reciprocally consistent (for example, in the case of functions that can be aggregated, the less detailed function must coincide with the aggregation of the more detailed one).

3.16 Composition of functions

The third type of transformation to be highlighted,[173] consists in the construction of composite functions.

Given a space "X" with "n" dimensions (x_1, ……., x_n) and a space "Y" with "m" dimensions (y_1, ……., y_m), let the "m" functions be defined in "X" as:

$$y_i = f_i (x_1, \ldots\ldots, x_n) \text{ with } i = 1, \ldots\ldots, m$$

and the composite function defined in "Y" as:

$$z = g [f_1 (x_1, \ldots\ldots, x_n) \ldots\ldots f_m (x_1, \ldots\ldots, x_n)] = h ((x_1, \ldots\ldots, x_n).$$

Assuming the "y_i" as operand functions, the "g" as the composition function and the "h" as the result function, the composition of functions is equivalent to the application of the operator "g" to the operand functions.

Many operators of this type can be defined, also by varying the parameter "m" to obtain operators that are unary (m=1), binary (m=2), and so on. In turn, operators can be combined in expressions to generate more complex operators.

For example, the ratio (i.e. the binary operator $g_1 = f_1/f_2$) between income (f_1) and the cardinality (f_2) gives the mean income (h_1). Alternatively, the logarithm (i.e. the unary operator $g_2 = \log(f_1)$) of income (f_1) gives the log income (h_2).

Composition operators normally presuppose that the operand functions be characterized by a common domain and compose the functions occurrence by occurrence, for the same element of the domain.

It is possible to define an algebra of this type not only in the space of the functions but also in the space of the variables, on the assumption that it can be applied only when the variable is dependent and only if the functions satisfy the applicability conditions of the operator (e.g. if they have a common domain).

[173] It needs to be remembered that in practice innumerable other types of transformations are defined and used.

3.17 Conclusions

The leitmotiv of this chapter is the notion of statistical function and, in particular, its intensional form. Ultimately, the component of the statistical system that brings together these intensional forms is the dictionary of statistical data.

The definitions are diversified. They specify not only the name and significance of the function but also the variables it is made up of, the related roles and sets of admissible values. It may also be necessary, in order to complete the definition of the domain, to specify the compatibility or incompatibility rules for the admissible values of different variables.

Redundant definitions can nonetheless be avoided. In fact, the function inherits the properties and relationships that are established between the concepts to be referred to, which are defined in the dictionary of concepts (e.g. the relationships of set-type aggregation between events, the definition domains of the variables, and the compatibility or incompatibility rules valid in general between spaces of events or between variables).

In effect, the primary characteristic of the statistical function, i.e. that of describing not single objects but groups, is already inherent in the significance and algebraic structure of the spaces of events (Boolean algebra), which is univocally defined in the dictionary of concepts.

The element linking the intensional and extensional forms is the knowledge domain of the function.

Moreover, according to the needs of a particular dictionary, the intensional definition of the function can be supplemented with suitable descriptive attributes and informal descriptions (e.g. methodological notes).

On the other hand, the correspondence between statistical data and algebraic functions makes the highly developed theoretical toolkit of algebra and mathematics available for the formal expression of the transformations to which a statistical datum can be subjected.

Consequently, the transformations themselves (or at least an important part of them) can be represented in the dictionary of data by means of a formal language.

It is thus but a short step, at least at the conceptual level, to render the dictionary “active” and to achieve the advantages, already amply demonstrated, of a similar architecture.

BIBLIOGRAFY

[1] Abrial J.R. "Data Semantics" in "Data Base Management" -Amsterdam, North Holland - 1974.

[2] Banca d'Italia "Le statistiche monetarie e finanziarie della Banca d'Italia" - March 1994.

[3] Banca d'Italia "L'informazione statistica nell'attività della Banca centrale", edited by the "Comitato per le statistiche creditizie e finanziarie", coordinator C. Conigliani, Tematiche istituzionali - October 1996.

[4] Banca d'Italia, Servizi ESI e ISC "Amministrazione Dati - Modellazione Concettuale - Manuale Operativo" - November 1992.

[5] Banca d'Italia, Servizio ISC "Progetto ASA - Specifiche Funzionali - Allegato C - Linguaggio per il calcolo delle variabili" - v.2.3 August 1995.

[6] Batini C., De Petra G, Lenzerini M., Snatucci G. "La progettazione concettuale dei Dati", Angeli - 1986.

[7] Batini C., Fortunato E. "Il progetto dei dati statistici", ISTAT - 1986.

[8] Batini C. "Struttura e progetto di un dizionario dati utente per la Banca d'Italia" - 1989.

[9] Batini C., Di Battista G. "A methodology for conceptual design of statistical databases" in "Inform. Systems", vol.17 n.3 - 1989.

[10] Bruschi G. "Basi di dati relazionali in Banca d'Italia" in "Office Automation" n.11 - 1986.

[11] Ciampi C.A. "L'informazione economica in Italia: problemi e prospettive", Roma - 15 Mar. 1984.

[12] Ciampi C.A. "La statistica nell'attività della Banca d'Italia" – speech at the "Università degli Studi di Roma La Sapienza" - in "Bollettino economico" of the Banca d'Italia n. 20 - 1993.

[13] Codd E.F. "A relational model of data for large shared data banks" in Commun. Association for Computing Machinery 13, pp.377-387 - 1970.

[14] Conigliani C. "La fabbrica delle statistiche monetarie e finanziarie", in "Bancaria", year 48 n.2 February 1992.

[15] Conigliani C. "La Centrale dei rischi", Banca d'Italia - Tematiche istituzionali, May 1995.

[16] De Marco M., Bruschi G., Manna E., Giustiniani G., Rossignoli C. "Sistemi elaborativi ed elaboratori elettronici" - ed. Il Mulino, 1987.

[17] Di Battista G., Batini C. "Design of statistical databases: a metodology for the conceptual step" in "Inform.System" Vol.13, n.4, pp.407-422 - 1988.

[18] EUROSTAT; EBES/EG6 - GESMES 95 Guidance to Users.

[19] EUROSTAT; EBES/EG6 - GESMES 95 Reference Guide.

[20] EUROSTAT; EBES/EG6 - GESMES/ECOSER Guidance to Users -January 1996.

[21] EUROSTAT; EBES/EG6 - GESMES/ECOSER Reference Guide - January 1996.

[22] Fazio A. "Statistica del credito", Associazione Bancaria Italiana, Roma, 19 June 1991.

[23] Johnson R. "Modelling Summary Data", in "Int. Conference on Management of Data", Association for Computing Machinery - SIGMOD - 1979.

[24] Langefors B. "Some approaches to the theory of information systems", BIT3, pp.229-254 - 1963.

[25] Langefors B. "Theoretical Analysis of Information Systems", Studentlitteratur, Lund - 1966.

[26] Langefors B. "Information systems", proc. International Federation for Information Processing, pp.937-945 - Amsterdam - North Holland - 1974.

[27] Langefors B. "Infological models and information user views", Inform. Syst. - 1980.

[28] Maggiolini M. "I processi di Analisi Dinamica su basi dati multidimensionali: le nuove opportunità sul mercato dei prodotti software", Banca d'Italia, Servizio ISC - September 1994.

[29] Magnani P. "Progettazione di un Data Base: I dati e i concetti fondamentali", in "Sistemi e Automazione" n.256 -1985.

[30] Malmborg E. "On the semantics of Aggregated Data", proc. of 3rd Int. Workshop on Statistical and Scientific Database Management, ed. Eurostat, Luxembourg - july 1986.

[31] Malmborg E., Lisagor L. "Implementing a Statistical Meta-Information System", in Eurostat Conference on Statistical Meta Information, Luxembourg - feb. 1993.

[32] Malmborg E., Sundgren B. "Integration of statistical information system - theory and practice", Eurostat Statistics, Telematic Networks & EDI Working Group 3rd meeting - dec.1995.

[33] Naddeo A. "Statistica di Base" ed. Kappa, II ed. - 1982.

[34] Padoa Schioppa T. "Le segnalazioni statistiche rivolte alla Banca d'Italia come fondamento della gestione delle aziende di credito", Associazione italiana per la pianificazione ed il controllo di gestione in banca e nelle istituzioni finanziarie, Roma, 12 Feb. 1993.

[35] Piccolo D., Vitale C. "Metodi statistici per l'analisi economica" - ed. Il Mulino, II ed., 1984 - pp. 261-262 .

[36] Rafanelli M., Ricci F. "Proposal of a logical model for statistical databases" in "Proc. of the 2nd Int. Workshop on Statistical Database Management" - Los Angeles - 1983.

[37] Raucci F. "Il riutilizzo del software nello sviluppo delle applicazioni: un'esperienza in Banca d'Italia" in "Sistemi Software" - October 1992.

[38] Serafini R., Lopez G., Milani P., Del Vecchio V. "Comments on BIS Questionnaire to be attached to the response", Bank of Italy Research and Statistical Service Departments, Sept.1996.

[39] Shoshani A., Chan P. "Subject: a directory driven system for organizing and access large statistical databases", in "Proc. of the 2nd Intern. Conference on Very Large Data Base", pp. 553-563 - 1980.

[40] Statistical Commission and Economic Commission for Europe - Standardization of Statistical Indicators and Metadata, Conference of European Statisticians - June 1995.

[41] Su S.Y.W. "SAM*: a semantic association model for corporate and scientific statistical databases", in "Inf.Sci.29" pp.151-199 - 1983.

[42] Sundgren B. "An Infological Approach to Data Bases" Statistic Sweden - 1973.

[43] Sundgren B. "Theory of Data Bases", Mason/Charter - New York - 1975.

[44] Sundgren B. "What metainformation should accompany statistical macrodata?" - Statistic Sweden R&D Report -1991:9.

[45] Sundgren B. "Statistical Metainformation and Metainformation Systems" - Statistic Sweden R&D Report - 1991:11.

[46] Sundgren B. "Organizing the Metainformation Systems of a Statistical Office" - Statistic Sweden R&D Report - 1992:10.

[47] Sundgren B. "Guidelines on the Design and Implementation of Statistical Metainformation Systems" - Statistic Sweden R&D Report - 1993:4.

[48] Tsichritzis D.C., Lochovsky F.H. "Data Models" -Prentice-Hall - 1982.

[49] United Nations Economic Commission for Europe "Statistical Journal" Vol.10, n.2, 1993 (special issue on statistical metainformation systems).

[50] United Nations Statistical Commission and Economic Commission for Europe "Guidelines for the Modelling of Statistical Data and Metadata", Conference of European Statisticians Methodological Material - United Nations -N.Y. and Geneva - 1995.