Data in the Natural and Exact Sciences and in the Social Sciences; a Comparative Study

by Eliahu Hoffmann[1]
National Center of Scientific and Technological Information, Tel-Aviv , Israel

Introduction

A comparison of data in the social sciences with those in the natural and exact sciences seems at first completely out of order. The nature of the data in both areas of science is so different that, content wise, they certainly cannot be compared. However, if we approach the comparison from a technical operational point of view we may try and classify them and compare.

1. the types of data that we find in both areas of science,
2. the treatment that we accord these data in both areas of science.

Such a comparison may be possible and may furthermore yield deepened insight and understanding in the meaning of the data. In addition, operational technical experience in data handling may be exchanged eventually between the two areas of science and this in turn may improve methods of data handling in both of them.

Types of data

In the natural and exact sciences three basic types of data can be discerned.

1. Fundamental constants and permanent data,
2. Time and space (or place) dependent data,
3. Factual data.

1. Fundamental constants are constants which have been measured directly or indirectly presumably "once and for all". They appear in numerous calculations and therefore their precise value is of great importance. If their values have been established a long time ago it may happen that, with better measuring instruments and methods they are being slightly corrected or improved. These corrections usually touch some of the last digits and thus make them more precise. By and large though, their values do not change anymore. A few examples may be quoted such as: The speed of light (299,792.8 km/sec or 300,000 km/sec)[2] ; Planck's constant (h) which represents the elementary quantum of action in physics[3] ; and even the standard length of the meter (0.0000001 of the earth's meridian quadrant at sea level)[4] .

Apart from these, there exist numerous series of permanent constants which have been measured and verified and are valid ever since. As examples we have: The values of the atomic weights of the elements; the infra-red (I.R.) absorption spectra of chemical compounds. Each compound, according to the structure, exhibits a series of absorption lines of specific wave lengths of the infra-red spectrum. These are determined by the structural characteristics of the compounds and serve as analytical tools to identify the compounds and their structures. Similar data series are: the solubility data i.e. the maximum quantity of material soluble in a certain volume of liquid. This quantity usually rises with the rise of temperature of the liquid and is constant for every temperature of the liquid.

2. Time and space (or place) dependent data are only true and valid for the place and time at which they were measured and describe a certain numerical situation at that place and time. Intensity measurements of incident sunlight may be taken as an example. These are usually high in summer and low in winter, higher in the equatorial regions and lower in the northern and southern regions of the earth and strongest at noon and zero at night. In addition, ecological data belong to this type i.e. the concentration of carbon dioxide in the air is high in urban industrial areas and low in rural areas. Data on rainfall are of a similar nature i.e. dependent on the time and the place when and where it occurs.

3. Factual data are essentially descriptions of properties of substances and/or processes. Among them can be named the properties of materials, whether they are hard or brittle, elastic or rigid, whether they are heat and/or electricity conductors or isolators etc. The taxonomic categories of classification of living organism belong to this type as well as elaborated DNA sequences in genetic material.

It seems that in the social sciences similar types of data can be recognized.

1. Constant data here too are data the validity of which is permanent and not dependent on - or restricted to - a specific time and place. (Though the data themselves may represent either dates or places). Personal records

for instance may be cited as an example. Thus social security numbers, identity card numbers, dates of birth, dates of immigration/emigration; educational records such as dates of awards of degrees etc. are all essentially constant data.

2. Most of the data in the social sciences however seem to be of the time and space dependent type. Census data, election results, manpower surveys, import-export statistics, fiscal and financial data, all these relate specifically to a certain time or time - period and to a certain place - city, country, continent etc.

3. As for factual data in the social sciences; historical and archival records, laws and regulations etc. can be considered to fall into this category.

Thus in spite of the fact that the data in the social sciences are very different in character from those of the natural and exact sciences, the same type of classification and characterization can apparently be applied to both.

Treatment of data

In the natural and exact sciences this consists essentially of 5 distinct steps:

A.) collection, B.) evaluation, C.) organization, D.) storage, E.) retrieval.

A.) Collection: Data in the natural and exact sciences are usually "buried" in textual research papers and have to be traced and pulled out. Thus, properties of materials or spectroscopic data, for instance, even for one substance only, may be found in a number of different papers since they are recorded also for analysis and identification purposes. If a list of data on one specific property for a large number of compounds is to be prepared, the search for them will extend through an even larger collection of research publications.

In the social sciences, by comparison, data re usually either collected and recorded routinely for well defined purposes. For instance, car license and ownership are registered, *inter alia*, for statistics of vehicle distribution; export - import data are used for balance of trade calculations; issue of identity cards serve population statistics and passport control shows the movement of people. In addition, they can also be collected by means of special projects i.e. surveys etc. In both cases the relevant data are obtained, *a priori*, in a more centralized fashion.

B) Evaluation. Since data in the natural and exact sciences are collected from a variety of sources, their evaluation is of great importance. From a systemic point of view one can recognize here the following 4 aspects to be checked.

B1) Conditions of measurement: These include the type of instrument used and the calibration of this equipment, the type of experiments performed and the methods by which they were measured. Were the correct, the relevant parameters measured? Were standardized measuring units used? Are the results reproducible? Have the experiments been repeated and how often?

B2) Methods of calculations: What kind of approximations have been used and what margins of error have been specified? Are both of them compatible with the conditions of the experiment and the precision of the measurement?

B3) Methods of presentation: Have the results been presented in a uniform fashion? What scales have been used, logarithmic, arithmetic? Can the data be read precisely from the graphs or tables used?

B4) Reasonability of the results: These should be judged in comparison with similar experiments and measurements. In chemistry, for instance, some physical data depend on the structure of the compounds. Some specific structural groups exhibit distinct and unique absorption in the infra-red spectrum and the appearance of such an absorption is taken as proof that the compound has indeed this structure.

Evaluation of data in the social sciences is apparently somewhat different. Conditions of measurement (B1) are presumably fixed in advance, either by routinely collecting specific parameters, like identity card numbers, or by specifying the parameters to be collected via surveys, by means of interviews and questionnaires. Recently, however, it has been recognized that surveys carry with them possibilities of error due to the "cognitive processes that respondents are required to exercise"[5] i.e. to interpret the question. These errors may be more prevalent with questions for information than the queries for numbers, but nevertheless they should probably not be overlooked. Methods of calculation (B2) like approximations and margins of error can also be specified and/or decided upon in advance and depend *inter alia* on the sample size and the response to the questionnaire. Methods of presentation (B3) can be chosen freely so that the essence of the results and their interpretation are emphasized and finally the reasonability of results (B4) depends *inter alia* on the tracing and elimination of technical and systematic errors.

C.) Organization: The organization of data is crucial for their efficient retrieval and proper utilization and in general can provide insight into their significance. Listings of data in the natural and exact sciences are usually presented in tables which are collated in Handbooks. In chemistry, for instance, properties of compounds - like boiling points or melting points - can be arranged in tabular form according to the names of the compounds and in alphabetic order or, according to their molecular weight; in this case from the lightest to the heaviest compound. This order would of course be different from the previous one. In most cases, the alphabetic order according to the names is sufficient, preferred and the only one used. A number of physical constants - and not only one - are usually tabulated alongside the name of the compound. Updating, however, must be provided for the addition of new compounds and their properties. The organization of data in the social sciences seems to be more complex and

multifaceted. In the natural and exact sciences orders and arrangements according to a single parameter, normally a name, are quite sufficient. In the social sciences, however, the ordering and arrangements according to specific groupings are very important because it is by means of these rearrangements that the insight, the meaning of the data become evident. Population data, for instance, must ultimately be arranged according to: a) Age groups; b) Communities; c) Areas of population; d) Professional groups; e) Groups of income etc. if one wants to understand their significance. On the other hand, there is here no need for updating because the data describe a certain situation in space and time which cannot be repeated at will. New updated data shall describe a different - time and space - situation.

D. Storage: Data in the natural and exact sciences are stored in printed form in Handbooks. They have to be characterized by specific terms, key words, identifiers etc. for tracing and retrieval by means of indexes. The same holds true for their storage in computerized form; a number has to have a computer address, a flag, an identifier etc. Basically these requirements for storage are identical in both areas of science, simply because they are a function of the operational methods of information storage (and retrieval) and are essentially independent of the subject matter at hand.

E. Retrieval: Methods of retrieval of data too are identical for both areas of science. For the printed form the indexes are used and for the computerized database the retrieval and search programs perform the task. They are strictly methods of information processing, independent of subject matter or content.

Similarities and differences

In the previous paragraphs it was shown that data in the two areas of science can indeed be compared. The types of data are categorized in identical fashion and their treatment exhibits considerable similarities. These results would seem surprising in view of the basic differences in the subject matter of the two areas. One is indeed tempted to ask: "If they are so similar, how do they differ?"

It appears that the difference lies in the approach to, and the emphasis on data which prevail in the natural and exact sciences. This in turn may be due to the following two factors:

1. The natural and exact sciences have more permanently valid data than the social sciences and that from the early beginning of their development. Indeed, measurements carried out in chemical experiments one hundred years ago are usually valid today and may be repeated with the same results. The time and place dependent data which constitute the majority of the data in the social sciences are of more recent origin in the natural and exact sciences, to mention the environmental and ecological data as an example.

2. The quantities of the available permanent data are simply "mind boggling"! In chemistry, alone, for instance, at least 6.5 million distinct compounds are known today. Assume that each compound has at least 10 single constants and 1-2 series of permanent data (infra-red data, solubility data etc.) and you have hundreds of millions of numerical data.

For this reason the proper treatment of data in the natural and exact sciences, though not an aim in itself, was always considered of paramount importance. So much so that within the International Council of Scientific Unions, the umbrella organization for all the scientific international societies, a special scientific international, interdisciplinary organization was established. This organization, called "The Committee for Data in Science and Technology" - CODATA, is concerned with the theory and practice of the proper methods for data handling. ❑

[1]Presented at the IFDO/IASSIST 89 Conference held in Jerusalem, Israel, May 15-18, 1989

[2]Parker, Sybil F., Editor in Chief, McGraw-Hill Encyclopedia of Science and Technology. 5th ed., McGraw-Hill Book Co., New York, N.Y., 7:683, 1982.

[3]*ibid.* 10:368

[4]*ibid.* 8:486

[5]Fienberg, Stephen F., Tanur, Judith M., "Combining Cognitive and Statistical Approaches to Survey Design", Science 243 (4894):1017-22, 24 February, 1989.