Chapter Two
On Data, its Quality and Pedigree

Jorge Luis Romeu
IIT Research Institute
Rome, NY 13440

June 7, 1999

Executive Summary

In this chapter we discuss the origins, definitions and ways of creating and identifying good quality materials data. We also discuss the process and characteristics of generating metadata and providing its pedigree. We also discuss the process and importance of assessing data and illustrate it with a real life example. Finally, we discuss statistical problems in the generation of data (i.e. sampling) for analysis as well as other important problems involved in sampling and in statistical experimentation. The references for this chapter are very specific to this area and, as done in chapter one, they are provided at the end of the chapter.

Introduction

Data are scientific or technical measurements represented by numbers or other means [1]. Its importance in scientific and engineering work can never be stressed enough. For example, the soul and heart of statistical analysis is the data. It is well known that bad data induces the so-called *GIGO* model: Garbage In, Garbage Out. In plain English, no statistical procedure will yield good results if the data used with it is bad.

In a similar way, data is essential for the materials engineer who is designing, say, an airplane part, during the process of selecting the right components. Bad data may result in a poor materials selection that may cause very serious manufacturing consequences. But telling apart good from bad data is not immediately nor easily done. For, at first glance much of the data looks alike, even when they may be completely different in more than one way –especially in their quality! The objective of this chapter is to provide some background to the materials engineer, in the direction of recognizing good data and setting it apart from bad data. To this effect, significant number of materials data analysis references particular to this area have been consulted and are given at the end of this chapter. These chapter references are distinct from the general ones in the appendix.

Broadly speaking, **good data** refers to accurate, complete and trustworthy ones, which can be easily accessed by the materials engineer from a reliable source, such as a handbook or database. Good quality data has, first, been carefully collected by a serious organization that follows strict quality control guidelines. This means that the collecting organization has carefully reviewed the experimental and test procedures of the data originators, checked it for consistency and registered this (**metadata**) ancillary information jointly with the analysis data itself. Then, the reporting organization will also check the analysis results via some internal and external procedures, known as data validation and certification. Finally, they will place the entire data information package in an accessible medium, such as an electronic database, where they can easily be retrieved and used by interested practitioners.

In short, good data can be recognized because it shows a **good pedigree** –and bad data does not! Data pedigree is difficult to define but easy to instinctively understand. In the same way that one would look at a dog's pedigree, one can look at data pedigree. A dog may look very good, as it walks at a dog show. But one is only sure that it is worth paying a large sum for it when one verifies its pedigree: who its parents and grandparents were, and how many prizes they got, its health history, siblings, breeding organization, etc. Finally, that all this information is certified by a respectable dog breeding society, to be sure that it has not been tampered with or has been sloppily or hastily evaluated.

In a similar way, good data is recognized by the entire folder that describes who collected and analyzed it, and how was all this done. What were the procedures used to check the quality of the experimentation, measurements and processing of this data, who checked it and to what level. All *this information*, which *comprises* the **metadata**, is *assessed* and *provides* the **data pedigree**. We will illustrate these concepts further in this chapter, with a practical example. We will compare two data sets, one "good" and one "bad", and explain why they are so, based upon the discussion that follows.

Metadata serves several purposes. In addition to helping establish the data pedigree, metadata also allows the comparison of different data sets and helps adjust them to the technological advances that occur with time. It also provides the grounds to establish the range and circumstances of the analysis results and of the reporting organization.

For example, a given stress value is only valid for specific types of components and under specific environmental and working conditions –obtained from the metadata. This information is useful for design engineers. In addition, a statistical analysis of these stress values and the conditions under which they were obtained may provide indication of what factors (and in what magnitude) affect the material stress values. This information is useful for research engineers and scientists. Finally, we can combine the metadata with an assessment of the data providers' organization. Then, we can use these factors to establish, via statistical analysis, if they help to identify and differentiate "good" from "bad" data. Some examples of these uses of data are developed in chapters four and five.

To better understand the practical importance of using good data and of its associated metadata and pedigree, we drafted a short list of questions about the engineering uses of materials data: How do we generate good data? How do we request and use them? How does a trustworthy data bank accept and process them? What is the data accreditation process like? How does one justify the cost of generating quality data sets?

The rest of this chapter discusses the important problem of assessing **good data** and of investigating how it is established. It also deals with the closely associated concept (and the establishment) of **data pedigree**. We review some issues that arise when we attempt to answer the above questions about **quality** of the data and their pedigree. We performed an extensive literature search and used it as the unifying thread to discuss some of the requirements for good data generation, process, analysis, validation, accreditation and certification, as suggested by several experienced researchers in this field. We also discuss the problem of the associated **metadata** characteristics required for the appropriate application of the statistical procedures described in [2, 3]. We finalize with a discussion of the statistical aspects of data collection and their consequences in the process of generalization of the obtained data analysis results (inference).

With this approach we expect to achieve several goals. First, we expect to satisfactorily answer the important questions of why data, metadata and data pedigree are important issues. Then, to discuss what metadata should be preserved, when recording the data, to assess and complete it. We also expect to provide a comprehensive list of the ancillary information that, according to relevant experts in this field, comprises the most important metadata as well as specific examples on how to use and assess them. We also expect to review the roadmap regarding how expert organizations implement their validation and certification procedures and what criteria they consider most important for it. Finally, we provide, at the end of the chapter, a list of references of specialized articles that cover these important topics in detail, for further in-depth reading.

With all the above, we provide a short but comprehensive document that can be used by practicing materials engineers and scientists as a quick overview of the data problems or by a beginner to rapidly become acquainted with its salient features.

Data and Metadata

Several well known specialists have discussed materials data problems at length. We will now borrow and elaborate on their incisive discussions. Barrett [1] states that since materials data originate from tests developed under specific conditions, we need to record the corresponding *metadata*, or *data about the data*. Without this ancillary information, the experimental results will loose their contextual meaning. For example, the use of fatigue information is closely associated with the conditions in which fatigue occurred and with the related material specifications.

Examples of metadata information include characteristics of test materials, specimens used, experimental and test conditions, measurement and calibration procedures, readings of the experimental results, specific ambient conditions, etc. In addition, metadata (i.e. such ancillary information) is used to perform statistical analysis, to compare different samples, to establish smoothing curves, as well as in the process of validating the data.

Metadata is often missing or incomplete, creating a serious void in the data collection effort. An easy solution would be to collect and store everything about the data. But this creates even larger problems. For, we must also think about the ease of the information retrieval by potential users of electronic databases –if this information is ever to be used. In order to facilitate its retrieval, the storage of materials information has to be well planned, then implemented in such a way that it is easily and uniformly accessed. To this end, extensive standard formats have already been established (e.g. see [4]).

Barrett adds that metadata can also be used in assessing which data sets to pool together. For example, apparently similar data sets may have some specific difference (say an ambient condition, material thickness, temperature, etc.) that sets them apart. These data must be tested using say, ANOVA or the K-sample Anderson Darling test procedures, prior to pooling them together. They can be pooled only if the statistical tests do not reject the hypothesis that they come from the same population. We will see several illustrative numerical examples of these problems in chapter five.

Also, experimental techniques improve or change with time. New parameters are identified that affect test results. For example, new research may indicate that humidity or ambient temperature (not considered before) may affect the measurements of, say tensile strength. If taken into consideration, the variability of the model (and of the tensile measurements) can be reduced and the precision of the estimations (say the allowables) can therefore be increased. If metadata are available we may correct the original data for these new developments (again, see chapter five).

Finally, the ancillary information obtained from the metadata also provides the variables for regression and analysis of variance or covariance, among other statistical procedures. The functions obtained can then be used to correct or reclassify the data, as well as to fill in for data gaps, where necessary and feasible.

Kaufman [7] discusses the work on standards performed by the ASTM Committee E-49 for the Computerization of Materials Property Data. This work deals with the problems of facilitating data storage and retrieval. Kaufman presents the ASTM Committee list of materials descriptors and of guidelines for reporting test data. The list emphasizes the importance of a unique format for the identification of metals and metal alloys and of polymers. A standard data format for the computerization of test data and mechanical properties is necessary to make comparisons between data sets. For example, if it is experimentally determined that ambient temperature is a factor in, say tensile strength, then we cannot compare, at par, two data sets where one of them includes ambient temperature in the metadata and the other one doesn’t. Such comparisons are valid when all the relevant fields are obtained and compatible. This shows the importance of recuperating all the information requested, in standardized formats, in addition to just reporting test data.

Standardizing the information content again raises the problem of the evaluation of data for quality and reliability. This is another crucial issue for those working in the field of materials data generation as well as for those who use the data in their engineering design work. For, in the same way that collecting all available information about the data is not a solution, storing all available data sets is not one, either. We must perform a selection.

This problem has also been thoroughly studied. Kaufman [6] discusses data quality and reliability issues. He provides several lists of guidelines for subjective assessment, validation, analysis and certification of materials data, based on the ASTM Committee E-49.05 report on data quality. Kaufman discusses several data levels. The lowest level is that of unanalyzed (raw) data, then the analyzed individual results, then mathematically reduced, then validated, then evaluated and finally certified data. The precise definitions for these materials data classification levels are included in [6].

Kaufman also discusses standard guidelines for database management, regarding quality and reliability and emphasizing on identification of data sources, proof checking of the data, correcting errors and assessing user satisfaction. In [5], Kaufman also provides extensive guidelines for data evaluation. He classifies them into subjective assessment, validation, analysis and certification and gives lists of activities for each category. The fictitious example we developed in the first section of chapter five is based on these issues. For, it uses statistical procedures to investigate the association between objective quality factors and subjective, experience based, data source classification.

Barrett [1] discusses, in detail, the guidelines outlined by Kaufman [5] as well as the problem of quality assessment of data sets. Mixing good and bad data doesn't improve a data bank –on the contrary it lowers the quality of the mix. In particular, mixing good and bad data increases the data variability, which in turn lowers the accuracy of the derived allowables. Barrett states that data can be evaluated through a complete process, that starts with assessing the organization that generates it and ends with a comparison of the originated test results with well accepted and certified results.

According to Barrett, an organization that creates data can be evaluated through its experience, accountability, bias, calibration practices and management attitudes such as the separation between data generators and evaluators. For, to avoid conflicts of interest, an independent group should carry out the data validation, if such validation is done within the same organization. All these factors show the degree of preparation of the data gathering organization and its data generation activity performance level.

In addition the personnel, the raw data and the validation activities also carry a strong weight. The associated personnel, with its experience, qualifications and attitude, also contribute to the data quality and credibility. Raw data itself can be checked for accuracy, outliers, physical properties, etc, as will be illustrated when we discuss distributions, estimation and testing, in chapters three and four. The data validation activities include statistical procedures for assessing consistency with known physical laws, parameter values and apparent similarity. These procedures will be illustrated through the testing, regression and ANOVA examples developed in chapters four and five.

Barrett provides a well-defined set of activities for the validation team. He provides lists of guidelines for the validation process and for establishing data quality indicators. The most important guidelines are to work with plural teams that include members of uniformly high experience and ability, that base their decisions on true consensus and whose members work within the limits of their knowledge and experience. Barrett suggests avoiding inclusion of members of suspect reliability, experience or known bias. He also provides a glossary of terms concerning data, quality and their validation process.

Finally, Barrett states that *certification*, as opposed to validation, is *the recognition by* a *warranting authority*, of the quality of the data. These authorities have to be uniformly recognized and well established and should certify only for their area of expertise. Examples include committees of professional societies, official organizations, etc.

Types of Data and Databases

There are different types of materials data and databases, since there are different types of uses for them. Materials data are thus collected, processed and organized accordingly. Rumble [7] discusses how materials databases can be classified according to different schemes that include data, user, and application and access types. He states that materials information should flow from data generators (e.g. testers) to data users (e.g. handbooks) as flows a slow moving river. Rumble states that such information flow consists of the four stages of data generation, analysis, aggregation and reanalysis.

As the computer has become ubiquitous, more work is automated and performed through or with computers. Much of the materials testing found these days is done this way. The resulting data collection from materials test equipment is, hence, computerized. Rumble calls this computerized collection of original test results data, *laboratory notebook databases*. They can be computer searched, analyzed, updated and manipulated, among other functions. And they also contain very useful ancillary (meta) data.

Rumble then states that *report databases* are those that provide analysis results of test data. They may include sophisticated correlations, graphical comparisons, coefficients, parameters, etc. They can appear in the literature (journal articles and technical reports) or in handbooks. They serve several functions, including derivation of properties, extension of data domain and improved understanding. Rumble underlines the importance of the data analysis stage and of the need to preserve the results of these (intermediate) analysis procedures. Such types of materials statistical data analyses are the object of this SOAR and will be illustrated and discussed in the following chapters.

Handbook databases, continues Rumble, compile data and other results into collections (e.g. [2, 3]) and constitute the data source of first choice. Not too many of them exist and their need is strong. Materials organizations such as AMPTIAC foster the creation and development of good materials databases. The present SOAR is part of such effort, since creating good materials data bases requires a clear understanding of the statistical procedures employed in deriving the materials properties included in them.

Rumble then classifies data targeted to specific applications, as *applications databases*. These are derived for convenience or for the quality of their data and built for solving specific problems These may be custom built, for some specific project, but they are usually not maintained nor updated beyond the life of such specialized work.

Rumble also discusses the classification of databases by user groups and presents tables of such uses. Data base uses include the calculation and evaluation of materials properties, the design, development, selection and performance evaluation of materials, failure analysis and product information. Databases can also be classified into personal, group, institutional, collegial or public, according to whom are their users and what kind of organization they come from.

Rumble concludes by discussing the problems of moving databases between types, of transferring data between them and of planning, retrofitting, maintaining, operating and completing metadata information, all of which is very expensive and time consuming, but necessary. For example, a personal database may be acquired by an organization that publishes a handbook and wants to include this institutional database in it. However, before such inclusion is possible, certain activities need to take place.

There are also several widely accepted databases, in addition to the already mentioned ones in handbooks [2 and 3]. They have been developed by other well-known and respected organizations and groups of users. Among them is the VAMAS (Versailles Project for Advanced Materials and Standards) group, described by Reynard [8]. Some of VAMAS technical working areas include: wear test methods, surface chemical analysis, ceramics, polymer blends and composites, bioengineering materials, weld characteristics, creep crack growth and low-cycle fatigue. VAMAS also has a data base certification process in place and a set of well-defined database standards.

The CODATA (International Council of Scientific Unions Committee on Data for Science and Technology) is another data gathering group, broadly representative of the industrialized nations. Barrett [9] describes the CODATA organization and its aims to assist the materials data base managers and to provide them with information on cost-benefits, standards, guidelines and terminology.

Other authors also describe the work of international database groups. For example, Kaufman [10] discusses the MPD (National Materials Property Data Network, Inc.) effort, a pilot network from Stanford University. Kozolov [11] describes work done in the COMECON Standards Reference Data (SRD) system, which performed work in the former USSR and Eastern Block. Nishima et al. [12] describe such data base work in Japan and Lu and Fan [13] describe the activities on materials databases in China.

Data Accreditation

Collecting good data and avoiding bad ones is of paramount importance. To emphasize this, we revisit some of the activities performed during Data Accreditation, the process that leads to providing the users with assurances about the quality of the data.

Munro and Chen [14] present a complete methodology for data evaluation. They divide the data into seven increasing classes of (un)acceptability level: unevaluated, research (preliminary and work in progress), typical (from surveys), commercial (manufacturer’s), evaluated (basic acceptance), validated (confirmed via correlations and models) and certified (standard references). Then, they provide a general data evaluation procedure.

If materials are not well specified, data is classified as *unacceptable*. If the measurement methods are not described and one is dealing with manufacturer’s data, it is classified as *commercial*. If it is survey data it is classified as *typical*. If it is subsidiary data it is classified as *unevaluated*. If none of the above, data is also classified as *unevaluated*.

If the data provides (or is checked against) standard reference values, it is classified as *certified*. Otherwise, if correlation or models have been applied, it is classified as *validated*. If data is checked by independent values, it is classified as *evaluated*. If the data is not checked but real properties are provided, the data is also classified as *evaluated*. If peer reviewed and part of an interim report, the data are classified as *research in progress*. Otherwise, if results are incompatible with materials, data may have to be reassessed and reclassified as either *evaluated or unacceptable*.

Munro and Chen give precise definitions of these classifications and also discuss the activities involved in working with them. In addition, the authors provide specific examples of applications of analytical, statistical and graphical methods to the validation of the data. Some of these methods will also be discussed and illustrated in this SOAR.

Kaufman [5] also discusses data evaluation and analysis. He provides a short list of activities and methods to be employed in the evaluation of different types of data and data organizations. These include the activities performed by the owner/maintainer of data sources, in the appraisal of individual raw data records and data sets, in the analysis and derivation of material properties and in the characterization of data sources.

Regarding the data owner, Kaufman suggests looking into the experience of the organization, of the personnel involved and of their quality control procedures, among other features. Regarding the raw data record, he suggests the assessment of the testing organization, of the completeness of material descriptions, of the completeness of the test methods descriptions, of the test data itself and of the validation procedures employed.

Regarding the analysis of properties, Kaufman suggests using graphical and statistical procedures and parametric modeling. He also suggests performing comparisons with other data sets and sources. Finally, and regarding data sources, Kaufman characterizes them by type of data, test methods, traceability, degree of evaluation, periodicity of updating process and supporting index.

Kaufman [6] again discusses quality and reliability issues of materials databases, as well as the ASTM Committee E-49 criteria. Standardization of the information is basic, because it allows uniform and universal access to it. Standardization is obtained through uniform fields in the database. The recommended field content descriptions include database name or acronym, full title, name of producer, address, telephone, types of data (e.g. raw, typical, statistically derived, evaluated), materials classes (e.g. polymers, ceramics, ferrous, non ferrous metals), property classes (e.g. mechanical, physical, electrical), independent variables (e.g. fabrication process, product form, thickness), testing variables (e.g. time, temperature), updating frequency (e.g. static, irregular, quarterly), evaluator name and organization, availability (e.g. public, private, free, fee) and media used (e.g. on line, hardcopy, compact disk).

Finally, Barrett [1] also comments about other database quality indicators. He discusses data presentation issues (e.g. accuracy), unit conversions and other data manipulations. Such issues are often taken for granted, but they are relevant to the values recorded.

Barrett provides a list of quantifiable quality indicators for assessing data records or databases. These indicators are grouped into data quality (e.g. source, statistical basis, evaluation status), database quality (e.g. completeness, support) and database operation (e.g. availability, access). Barrett states that this list of indicators may be regarded as a vector in a multidimensional space. Under this multivariate approach, database comparisons may be established by looking into each component.

An Illustrative Example of Data Comparison

To provide an illustrative example of application of the above data discussion, we have asked Mr. David Brumbaugh, of AMPTIAC to select and compare two data sets. One of them is "good" data, while the other is "bad" data. The technical reasons Mr. Brumbaugh has provided, for performing such data classifications, are as follows:

In this example, data on the yield strength of 1-inch annealed 4130 steel bar has been taken from two different sources. The first source [18] is the Aerospace Structural Metals Handbook (ASMH), and the second source is from an Internet web site that will remain anonymous for the purposes of this discussion. The first noticeable difference in the data is that there is large difference (15ksi) in the values reported for the yield strength (82ksi ASMH and 67ksi Webster). The second difference is that the data from the web site lacks essential metadata such as processing history, whereas it is given in the ASMH. For example, in the ASMH, processing history such as the fact that the material was cold worked and then annealed at a temperature of 1550 Fahrenheit was given. No annealing temperature or other processing history was noted in the data from the web site. The third and last major difference in the data can be seen in the sources cited for the data. The format in which the references are given for the data taken from the Webster lead to uncertainty in determining from which reference source the yield strength was obtained. For instance, it is unclear whether the data was obtained from a questionable source such as vendor information or from a very credible source such as the ASM Metals Handbook. The ASMH on the other hand, is very clear in identifying the two sources from which the yield strength data was obtained. Because of the completeness of the data obtained from the ASMH, it would in this example be considered the good source, whereas the web site data would be considered the bad source.

Uses and Cost of Good Data

So far, we have discussed materials data, their quality and their pedigree. And we have provided an illustrative example of how to detect and classify a data set. Obtaining good data however, does not come free: it entails a cost. On the other hand, an (economic) benefit is also derived (directly or indirectly) from its uses in engineering design. In this section, we discuss issues of good data uses and their corresponding costs and benefits.

Newley [15] presents a case study example of the integration of materials information into engineering design. He describes the evolution of an information system from the initial recognition of its need. Some of the advantages of creating materials information systems are derived from their function of providing a central source. Information systems provide a source of best *available* data, that designers and analysts can use in their work. Information systems also provide a source of preferred materials and processes as well as of experience gained in manufacturing them. They provide the fact that data used is traceable and the possibility that one is now able to assess which information is most valuable.

Newley states that materials information is required throughout the life of a design. He then provides an information flow process for the five stages into which he divides the engineering design process. These stages are R&D, product scheme, detail drawing, production qualification and in-service product report. In all of them, the materials information process has a valid, distinct and useful input.

Newley then provides a list of technical factors affecting materials selection that can be included in such life cycle information flow. They include specific materials properties (e.g. fracture mechanics, fatigue, strength and ductility), compatibility (e.g. corrosion, wear, thermal mismatch) and manufacturing (e.g. availability, cost, machinability, inspection, formability). Newley finishes by describing the requirements of materials data information systems and outlining its data structure organization.

An example of the use of a software tool for materials data analysis is given by Zhou et al. [16]. The software includes six advanced statistical analysis procedures. They are nonlinear mapping, principal components, stepwise discriminant analysis, discriminant analysis with constellation graph, hierarchical clustering analysis and stepwise regression. It also has an artificial neural network algorithm. Using such a software package for materials data analysis and property prediction presupposes the existence of a database with abundant and reliable data. These advanced quantitative statistical analyses can greatly enhance a design process. However, they are only possible when plenty of good, quantitative data are available and easily accessible.

The software described by Zhou et al. first takes a data set and pretreats it (analyzes it in a preliminary way) looking for relationships (in a similar way as we will do in chapters four and five). Then, it applies diverse multivariate statistical procedures, according to the features of the data set and the objectives of the study in question. The software then provides graphical and analytical results. Several engineering design case studies are presented, that demonstrate the various types of data analyses that can be performed with this software, when one has access to a large amount of good and reliable data.

However, good and reliable data does costs money to collect, validate, install, deliver and maintain. Barrett [17] discusses the many benefits and economic consequences of such materials databases. Barrett defines the current problem of justifying data collection as one of operating in a *market driven* economy. In these economic times, one is required to perform a cost-benefit analysis of the engineering information system and to show the real value added by it, to the design process. The problem with this approach is that the information activity hides its benefits quite well. And it is easier to show the losses incurred, by not using good information in the design process, than it is to show the gains obtained by using good information systems.

Regarding this situation, Barrett suggests that cost-benefits relationship should be uncoupled until benefits are better characterized and understood. He also suggests that different viewpoints on information benefits also be recognized. These viewpoints should include not only those from system developers, but also the viewpoints of users (of existing systems) as well as of potential users (of new systems under development) and also the viewpoints of those non-users who can influence the process (e.g. managers).

Barrett suggests and describes a new approach to the quantitative evaluation of benefits. This approach presupposes that economic benefits do exist and hence should reflect somewhere in the system. Therefore, he proposes that database functions and features be linked with tangible user benefits –some of which may not be readily identified or appreciated. Some examples of such functions and features include the speedy access (by engineers and designers) to electronic data and the organization and structure of the information and electronic communication (that saves valuable search time of technical personnel). These work times that have been saved could be quantified and presented as a tangible economic benefit of having an information system in place.

There are also examples of economic and social advantages perceived or sought by the user of a materials information system. They include reduced design cycle time, lower labor costs, lower material and capital costs, improved product quality and reliability and enhanced education and work interest for the information system user. All these factors are quantifiable in dollars and cents and constitute evident examples of socioeconomic gains, introduced by the use of databases and information systems. Finally, Barrett provides tables where both, benefits and functions, are linked together establishing which functions yield what advantages and vice versa.

Statistical Characteristics of Good Data.

Data is the life and blood of materials data analysis and of statistical analysis in general. In this section we will discuss some general characteristics of “good” statistical data, as it relates to our materials subject matter.

In statistics, when we talk about data we think of a sample. We also think of data as the problem information from which we will derive our conclusions. We want the sample to be representative of the population it comes from, in order to infer (generalize) the results obtained, back to the entire population. Samples that are not representative (of the entire population they are drawn from) do not allow these types of generalizations. The analysis results then apply only to this sample or at best, to the population subclass they represent.

In order to be “representative”, a sample should be drawn at random from the entire population and each sample point (or observation) should be independent. In this case we say that every observation or measurement (and by extension, the sample) is independent and identically distributed. In statistical literature this is denoted as i.i.d. Searching for this representativity condition, the handbooks [2 and 3] request that measurements are taken from several batches, which in turn should include several measurements. Having the necessary population representativity and independence, enables us to generalize the data analysis results (e.g. materials allowables) to the entire population (of materials) from which it was drawn and not only for the particular specimens in that batch (sample).

Randomness (e.g. occurring by chance) is the property that allows an observation to appear in the sample with the same probability with which it appears in the population. For example, a specimen with damaged surface will randomly appear in the sample (by chance) once every 1000 times (in the long run) if it is observed in the population with this same frequency. Independence is the property by which the presence (or absence) of a specific sample value has no influence in the presence (or absence) of any other sample value. For example, having drawn a specimen with damaged surface, has no bearing on the fact that the next sample specimen drawn also has to have (or cannot have) a damaged surface.

Some times total randomness and total independence, are theoretical conditions that we can only strive for. On the other hand, disregard of these two conditions (say, by performing experimentation on specially developed lab material, under specially developed testing conditions) destroys the necessary representativity that enables us to generalize the experimental results to the entire population (statistical inference). In such case, we could not state that the experimental results are valid for ordinary production material operating under ordinary working and environmental conditions.

In addition to being random, independent and identically distributed, “good” samples should be as large as feasible for two reasons. First, as sample sizes increase the uncertainty (variance) associated with the information (point estimators) they convey, decreases. Secondly, a fundamental statistical result (the Central Limit Theorem or CLT) that will be discussed in chapter four will apply only to large and “good” samples. If our sample is “good”, and large enough that we can apply the CLT, our statistical work is greatly facilitated and our inferences, greatly enhanced.

Therefore, the minimum size that a sample can have is two elements (in order to obtain a sample variance and thus, a measure of uncertainty of the estimations). In addition, if the sample has at least 30 observations, the CLT and all its nice statistical properties, apply.

Some Statistical Sampling Schemes and their Characteristics

We sample because we do not have the time, the money or the physical possibility to measure the entire population. However, we still need to obtain some information (estimations) from, or to experiment with (a part of) the population, and then to infer our analysis results back to the whole. A good example of the practical relationships, problems and dynamics of the sampling versus counting controversy is illustrated by the current Congressional debate about the Year 2000 Census (and the under count of certain population groups e.g. the homeless, minorities, illegal aliens, etc.).

Populations can be finite or infinite and samples may be drawn from them with or without replacement. This has an important effect in how the estimations, and essentially their measures of uncertainty (e.g. sample variance), are calculated.

If the population is finite (say, of size N) and homogeneous, then any member drawn at random has a probability of selection of 1/N. A sample is said to be drawn with replacement when every observation taken from the population is put back into it. Each sample element has, at least theoretically, the same probability of being drawn again. But if observations drawn are not replaced, then different probabilities of selection apply to each sample element. For example, the first sample element has probability of selection 1/N, the second sample element has 1/(N-1), the third, 1/(N-2), etc. If N is small, the differences in probability of selection between the first sample element drawn and the last one (of a sample of size n) will be of 1/N to 1/(N-n+1). This can be quite significant if the sample size n is large. These conditions affect both the sampling distribution of the statistic used as estimator as well as its variance. In addition, once an element is drawn (and then discarded) its probability of being drawn again is now zero. We will revisit the sampling problem in chapter four.

Also in sampling without replacement, if the random sample size n is not large (but the finite population size N is) we may neglect the above mentioned difference in probability of selection (between the first and last sample elements) for it will not be significant. Then, for all practical purposes we assume that the probability of selection is the same. We will assume likewise when the population size is infinite and we sample with or without replacement. A “litmus” test for assessing when the above situation holds is to look at the relation (1 – n/N). If it is close to unit, then we are probably OK. If it is not, then we want to consider other alternatives.

Another important sampling consideration, in addition to the population size, is the population homogeneity (or lack thereof). If the population is homogeneous, all members have similar characteristics (as, say, in the tensile strength of a material that has the same thickness). Then, a simple a random sampling scheme is usually adequate and things are greatly facilitated. On the other hand, if the population is heterogeneous, there will be subgroups with different characteristics (as, say, in a material that has several levels of thickness). Then we may have a stratified population with strata or classes of different sizes. In this case, simple random sampling may not provide the best point estimators, for their variance will be larger than that of the estimators obtained by implementing other more appropriate sampling schemes (e.g. stratified sampling).

In materials data analysis, for example, we take several batches from different production runs. The objective of this approach is precisely to obtain the most representative sample possible, in order to extend our analysis results to the entire population (e.g. the whole production process) and not to restrict the inferences to the few batches considered. However, in practice batches may differ (i.e. constitute population subclasses or strata). This may occur because of specific production process characteristics. For example, a batch at the start of the production run may be produced under a different mix, different temperature, different machine adjustment, etc. than a batch produced at the end of the run). This may also occur because different producers may have different characteristics. For example, one manufacturer may have some special equipment, a different supplier, diverse management or technical qualifications, etc.

For these reasons, we first analyze whether the samples (batches) are similar (i.e. the population is homogeneous) in order to assess whether we can pool them together. If the statistical tests reject the hypothesis that batches come from the same population, then we assume that the population is heterogeneous and we do not pool the samples together. The way in which we calculate sample estimations, say the allowables, will differ in both cases. For, we strive to obtain sample estimators (e.g. allowables) valid for the entire population in question and not only for the specific batches (sub classes) analyzed.

This situation also has bearing with the “random effects” model of ANOVA and regression, discussed in the handbooks [2, 3] and in the RECIPE program manual [19]. In the “fixed” effects model, inferences (generalization) are reduced to the finite population of the batches (groups) submitted to analyses or, at best, to the specific production runs that generated them. On the other hand, in the “random” effects model, the selected groups or batches and their generating processes, are themselves considered random representatives of all possible batches from all possible (similar) generating processes. The inference is drawn for this broader population, which is what the materials engineer is looking for. Elementary information on sampling can be found in any of the statistics textbooks referenced in the appendix. More advanced information can be found in [20].

A final word is due regarding the case of special studies performed on non-random samples. If (pilot) experimental research is undertaken under, say, special (laboratory) conditions, then two considerations are in order concerning their results. First, pilot experimental samples are usually not random and hence, their inferences should not be automatically generalized to an entire population (which probably lies outside of its inference range). Secondly, the inferential restrictions do not imply that the experimental results are lost or worthless.

On the contrary, these experimental results are extremely valuable for what they are: an initial or pilot research study. They provide several very useful pieces of information, among them an initial estimation of the variance of the measurement (random variable) of interest. Having such initial estimations are essential for designing subsequent studies and experiments, specifically in determining the sample size required to obtain further results within certain value ranges. We will revisit this important issue when we discuss hypothesis testing, in chapter four.

In any case, experimental results will be enhanced if the pilot studies are conducted under the most “representative” conditions. One way of achieving this occurs when the researcher selects the specimen types, operating conditions, environment, etc. according to his experience. The samples obtained with this approach are still subjective and will never surpass a randomly drawn sample. However, if the researcher’s judgement and experience are correct, they will be closer to representative conditions and will provide results closer to those obtained from a random sample. Some times, such experimental samples selected by the good judgement of an experienced researcher, are referred to as judgement sampling. They constitute a valid option in pilot studies, and yield very valuable contributions in the process of scientific research.

Conclusions

In this chapter, we have summarized the main issues regarding materials data. We have discussed their quality and their pedigree, as well as their relationships to the construction, maintenance and use of "good" materials databases for engineering design. Finally, we have provided an illustrative example of what a "good" database is and what it isn't, and why. In addition, we have discussed some important statistical issues regarding samples, sampling schemes, their characteristics, their limitations and their (statistical inference) consequences in the process of scientific research. Data is the analysis raw material and random sampling is its best source. This chapter, therefore, develops the insight for better understanding all the chapters that follow and for a more efficient implementation of the statistical procedures used in materials data analyses.

Bibliography

1. Barrett, A. J. (1993). "Data Evaluation, Validation and Quality". ASTM Manual on The Building of Materials Databases. Crystal H. Newton, Editor. ASTM Manual Series: MNL 19. Pages 53 to67.

2. MIL HDBK 5G. Metallic Materials and Elements for Aerospace Vehicle Structures. November 1994.

3. MIL HDBK 17. 1D. Composite Materials Handbook.

4. ASTM Standards on the Building of Materials Databases. ASTM Series (1993).

5. Kaufman, J. G. (1989). "Standards for Computerized Material Property Data –ASTM Committee E-49". Computerization and Networking of Materials Databases. Glazman and Rumble, Editors. ASTM STP 1017. Pages 7 to 22.

6. Kaufman, J. G. (1992). Computerization and Networking of Materials Databases: Third Volume. Barry and Reynard, Editors. ASTM STP 1140. Pages 64 to 83.

7. Rumble, J. R. (1993). "Types of Materials Databases". ASTM Manual on The Building of Materials Databases. Crystal H. Newton, Editor. ASTM Manual Series: MNL 19. Pages 27 to 33.

8. Reynard, K. W. (1989). "VAMAS Activities on Materials Data Banks". Computerization and Networking of Materials Databases. Glazman and Rumble, Editors. ASTM STP 1017. Pages 43 to 52.

9. Barrett, A. J. (1989). "CODATA Activities on Materials Data". Computerization and Networking of Materials Databases. Glazman and Rumble, Editors. ASTM STP 1017. Pages 89 to 106.

10. Kaufman, J. G. (1989). "The National materials Property Data network, Inc. – A Cooperative National Approach to Reliable Performance Data". Computerization and

Networking of Materials Databases. Glazman and Rumble, Editors. ASTM STP 1017. Pages 55 to 62.

11. Kozolov, A. (1991). “Materials and Substance Data Banks in COMECON Countries and in the USSR”. Computerization and Networking of Materials Databases: Second Volume. Kaufman and Glazman, Editors. ASTM STP 1106. Pages 7 to 16.

12. Nishijima, S, Y. Monma and M. Kanao (1989). “Japanese Progress in Materials Databases”. Computerization and Networking of Materials Databases. Glazman and Rumble, Editors. ASTM STP 1017. Pages 80 to 91.

13. Lu, Y. and S. Fan (1989). “Materials Data Activities in China”. Computerization and Networking of Materials Databases. Glazman and Rumble, Editors. ASTM STP 1017. Pages 75 to 79.

14. Munro, R. G. and H. Chen. “Data Evaluation Methodology for High-Temperature Superconductors. (1997). Computerization and Networking of Materials Databases. Nishijima and Suichi, Editors. ASTM STP 1311. Pages 198 to 210.

15. Newley, R. A. (1992). “The Integration of Materials Information into Engineering Design”. Computerization and Networking of Materials Databases. Barry and Reynard, Editors. ASTM STP 1140. Pages 192 to 205.

16. Zhou, J., X. Qian, J. Feng, S. Li, Z. Xu, L. Chen and Z. Gui. (1995). “A Software Tool for Material Data Analysis and Property Prediction: CASAC-ANA”. Computerization and Networking of Materials Databases. Sturrock and Begley, Editors. ASTM STP 1257. Pages 235 to 252.

17. Barrett, A. J. (1991). “The Benefits and Economic Consequences of Materials Property Databases”. Computerization and Networking of Materials Databases: Second Volume. Kaufman and Glazman, Editors. ASTM STP 1106. Pages 17 to 25.

18. Aerospace Structural Metals Handbook (1997). CINDAS/USAF.

19. Vangel, M. (1995). A User’s Guide to RECIPE. NIST.

20. Sukhatme, P.V, B. V. Sukhatme, S. Sukhatme and C. Asok (1984). Sampling Theory of Surveys Applications. Iowa State University Press (3rd. Edition).