Chapter Eight
Conclusion

Jorge Luis Romeu
IIT Research Institute
Rome, NY 13440

June 4, 1999

Executive Summary

In this last chapter we summarize what has been accomplished in the areas of statistical analysis of materials data. We also discuss possible topics for future work in the area of statistical training for materials engineers and scientists. Finally, we overview the data sets and software used in the case studies presented.

What Has Been Accomplished

As we explained in the Introduction chapter, the objective of this SOAR is to provide support and extension materials for better understanding the data analysis statistical procedures in handbooks [6, 7]. It can also be used as reading material by practitioners in statistical analysis, either as a methods refresher or as a discussion on the statistical thinking behind the procedures used in the mentioned handbooks.

We have covered several important topics. We have discussed univariate and bivariate statistical methods, for both, qualitative and quantitative variables. We have seen the most important and frequently used statistical distributions in materials data analysis. Among the continuous distributions, we have discussed the Normal, Lognormal and Weibull. Among the discrete distributions we have worked with the Binomial and the Discrete Uniform.

We have discussed the meaning, interpretation, use and estimation of distribution parameters and other performance measures of interest. These include measures of central tendency such as the mean, median and mode and measures of dispersion such as the variance, ranges and coefficient of variation. Finally, we have discussed other measures of location such as the percentiles, which are relevant to the study of the concepts of A and B basis allowables.

We have discussed and provided contextual examples for several sampling distributions of interest, such as the Student t and Fisher's F distributions, used in the comparisons of two or more samples. We have also discussed the Chi-Square distribution, useful in testing for association and for testing the variance of a sample. Finally, we have discussed (and provided examples for) the use of several non-parametric distributions, especially for the Anderson-Darling, Mann-Whitney and Kruskal-Wallis non-parametric tests for the comparisons of two or more samples.

We have discussed in detail (and have provided GoF examples of) the important problem of establishing the (underlying population) distribution of a sample. We have then discussed (and provided several graphical and analytical procedures for) detecting potential outliers in a data set. For, once a statistical distribution has been established, we want to know if all the data in the set conform to it. We have also discussed the serious problems and consequences of detecting and mechanically discarding outliers.

We have discussed in detail the implementation of confidence intervals and hypothesis tests, for certain distribution parameters of interest. We have also provided contextual examples and step-by-step procedures for developing them. In particular, we have discussed the derivation of large and small sample confidence intervals for the mean. We have also presented practical examples of the derivation of confidence bounds and tolerance bounds and intervals.

We have discussed in detail the implementation of hypothesis tests for the mean of one and two populations and the tests of association between two qualitative variables (contingency tables). We have also provided contextual examples and step-by-step procedures for developing them. We have provided interpretations for types I and II errors and their corresponding risks. We have discussed the problems of establishing the sample size for a pre-specified confidence interval and their statistical consequences.

We have discussed the problem of the implementation of one-way analysis of variance (ANOVA), the model assumptions and several graphical and analytical procedures to check their validity. We have discussed procedures for comparing the means of several samples and for establishing joint confidence intervals for their differences. For, these are required once the ANOVA procedure detects that the population means do differ.

We have discussed the problem of implementing simple linear regression and non-linear (quadratic and cubic) regressions. We have discussed the regression model assumptions and presented several graphical and analytical procedures to check their validity. We have also discussed the important problem of fitting several (linear and non-linear) regression models to the same data set and then selecting the one that best fits (or describes) the underlying problem structure. Finally, we have discussed, in detail, the dilemma of modeling the data instead of the problem, ever present in data analysis.

Finally, we have presented two complete chapters discussing case studies of real life data analyses: one in ANOVA and another one in regression, respectively. The data have been borrowed from the handbooks [6, 7] and the RECIPE materials data analysis program [5]. We have developed these case studies as we would have analyzed them for a research project. We have started with the EDA (exploratory data analysis) description of the data. We have proceeded to the formulation of conjectures (hypotheses) that would have then been tested. Finally, we have actually tested these hypotheses via the implementation of different statistical models, such as ANOVA, regression, t and F and non-parametric tests. We have checked, in each case, the validity of model or test assumptions. Finally, we have derived the corresponding statistical and problem context conclusions.

Data and Software

Two necessary ingredients for data analysis work are the data (raw material) and the analysis tools (statistical software). In the previous chapters we have thoroughly used and discussed both of them. We will now provide some other technical details about them.

At the start of this SOAR we dedicated an entire chapter to discuss data, its quality and its pedigree. We did this both, from the specialized point of view of the materials data analysis as well as from the purely statistical one. We included a long list of materials data literature sources, for the interested reader to pursue this topic further. We also discussed ways to check that the data used is reliable, via the materials science concepts of data pedigree, validation and accreditation. This says how important is data to us.

Then, we extensively used several real data sets, that we identified in the corresponding chapters. We present an annotated list in Appendices 2 and 3. For the SOAR readers convenience we also include text files of these data sets. With this, we hope to encourage the readers to redo the analyses and gain the practical experience.

To implement these analyses some statistical software is needed. We have tried to keep such need to a bare minimum by limiting our own use to two packages: Microsoft Office Excel and Minitab. Both can be used in the analyses. Many of these analyses were and can be readily implemented using Excel spreadsheet formats and its linear regression and graphical capabilities. Minitab is a specialized (statistical) software package widely used in statistical education. Minitab web page is http://www.minitab.com/.

We have used the student version of Minitab, which is independently available at a nominal cost. It also comes with many general statistics book. It is easy to learn and use and this author has utilized it for years, in his statistics college courses. Any other statistical software (e.g. SAS, S-Plus, SPSS, etc.) would also do. For, the procedures used in this SOAR are very general and are included in most statistics software packages. We have deliberately used both, Excel and Minitab, to make the point that these procedures are at the reach of almost any one who has access to a PC. No special endorsement should be inferred about either software.

Finally, there are several specialized materials data analysis statistical packages in the Web, of easy and free access,. The reader can find some of them in the NIST Statistical Engineering Division’s Web Page: http://www.itl.nist.gov/div898/. We have not used this specialized software here because the intent of this SOAR has been to discuss the implementation of statistical procedures. The use of specialized software in our analyses would have defeated such SOAR intent. In our next and final section, we propose several subjects for future or extension work. The development of training materials that discuss the access and use of such specialized statistical software, is one of them.

Future Work

We have accomplished several objectives in this SOAR. However, this is only the beginning of a long journey into the vast world of statistical data analysis. The interested traveler may want to find out what are some other possible topics that might follow.

So far, we have seen material on univariate and bivariate data analysis. When there are more than two pieces of information per observation (say k pieces) we have a k-variate (observation) random vector. Then, we deal with k-variate (in general multivariate) data analysis. The statistical problems discussed for the case of one and two dimensions are now magnified by the higher dimensionality of the problems.

However, the analysis possibilities also increase proportionately. For, now we can look at the relationships between the random vector components in multiple and very interesting ways. We start (as in univariate) by establishing the underlying multivariate distribution via multivariate GoF tests. After this distribution is established (and according to the specific research needs and objectives) we may perform discriminant or factor (classification) analyses, among several other candidate multivariate procedures.

In discriminant analysis we predefine the sub-populations. Then, the statistical procedure processes the multivariate data using as criteria the variables or components. If there is a real difference between the pre-established subgroups, then some of the discriminant function variables will result statistically significant. The resulting equation will be used to identify the sub-population to which each observation probably belongs to. For example, one may hypothesize that a given material is divided into two sub-populations: “good” and “poor” quality. Then, if this hypothesis holds, the discriminant analysis will establish which variables (say heat, temperature, humidity) actually “discriminate” between them. We can then use this equation to classify say, an incoming material, into one of these two sub-populations.

Factor analysis (and its related principal components procedure) looks at the internal correlation between the random vector variables and then regroups these components into “factors”. For example, there may be a data set (with, say ten variables) that can be regrouped into (ten orthogonal) “factors”. Of these factors, say the first two may describe 80% of the variation in the data. These two factors may even be identified with specific meaning or abstract materials qualities (say, “durability”).

Thus, factor analysis provides two useful advantages. First, it reduces the number of observation components with which we have to work (say from the ten initial variables to the two “factors” that describe 80% of the problem variation). Then, it provides a grouping scheme (if we select the first two or three orthogonal “factors” we may be able to separate the observations into cohesive groups, in a two or a three dimensional space). This scheme may help us better see, understand and study their common traits. We must always keep in mind that one of statistics main goals is to analyze a data set in order to determine how (factors, groups, variables, etc.) are alike and how they differ.

Such types of classification analyses help determine whether the observations can be separated into subgroups that have certain characteristics in common. An important practical difference between discriminant and factor analyses is that, in the former, the population subgroups are predefined by the analyst. Then, the statistical procedure assesses whether these groups actually differ, and in which variables. In factor analysis, the groups, if they exist, come out as a result of the analysis procedures themselves.

Another two useful topics for possible future work are general (multivariate) regression and ANOVA. As can be surmised by the reader, these are also two special cases of multivariate analyses. In this SOAR we have dealt with linear and quadratic regression and with one-way ANOVA. In the extension or future work, we would expand the analyses to more than two dimensions or factors. That is, we would regress one variable into several, or analyze one response on several factors, not just on one, as done here.

For example, we can model a response (say, tensile strength) as a function of several factors (say, thickness, temperature, age, etc.) and their first, second, etc. order interactions (say, the interaction of say, thickness with temperature, etc.). This yields the multiple regression case where the independent factors or predictors and the response, are all quantitative variables.

However, we can also deal with qualitative variables. For example, we can model the same response (tensile strength) but now as a function of several qualitative factors (say, manufacturer, batch, heat, etc.) and their first, second, etc. order interaction (say, the interaction of batch and manufacturer). This yields ANOVA, where we are dealing with multiple factors (two way, three way, ANOVA etc.) and possibly with their interaction.

In the same vein as above, we can explore Design of Experiments (DoE). Here we apply specialized analysis of variance models to scientific experimentation. The different statistical models or designs (say, fractional factorials, split plots, cross over, etc.) have been developed to accommodate special research situations. One example occurs when the experimenter wants to assess the effects of several factors on a response and tries to minimize the total sample, the number of experimental setting combinations used, etc.

Sampling is also an important subject to present. For, the sample (or data set) is the raw material from which all statistical results are derived. Experimenters would like to extend their results to the entire population and not to restrict them to the data analyzed. In this SOAR we have worked with completely randomized samples (i.e. where each subject or each combination of subjects has the same probability of being selected from their population). But there are populations where a different sampling design (say, stratified, conglomerate, systematic, etc.) may be used to the advantage of the analyst (in the sense of reducing the variance of the population parameter estimates).

In materials data analysis, testing is done on specimens drawn from the population of all possible materials of a certain type. Getting the smallest but most efficient and representative samples (in the sense of least estimator bias and variance) is of importance in order to be able to generalize the research findings to the entire population of such type of materials, while keeping sampling costs down.

Quality control is yet another subject of interest for material scientists. For, we want to have consistent (as opposed to variable) quality in the materials that we work with. Acceptance sampling plans, control charts, sequential and censored testing procedures as well as newer quality topics such as TQM, SPC, Six Sigma and others, may also be of interest to those working in this area.

Finally, statistical computing is one of the hottest topics, today, in technical schools and universities. The reason is that practically all statistical analyses are currently done using statistical software. Hence, software has become as complex as it has become powerful.

As we have already mentioned in this SOAR, there are excellent specialized computer programs and spread sheets that have been written for, and deal with, materials data analyses. Many of them can be easily and freely accessed via the Internet. Among these we find the RECIPE program, as well as Dataplot and Omnitab, which can be accessed via NIST's Engineering Division Web site: http://www.itl.nist.gov/div898/. Another useful data analysis program, STAT-17, can be accessed via the MIL HDBK 17 Web site: http://mil-17.udel.edu/. In addition, there are excellent commercial statistical packages such as Minitab, SAS, S-Plus, SPSS, etc. Materials data analysts may also want to learn to use such software, for the data analysis freedom they provide.

Some of the above topics can become stand-alone SOARs. Alternatively, they may be combined into a single one, as needed or requested by the materials data analysis community. The development of short courses, videos, distance learning materials, etc. with these topics may be another interesting possibility for those interested in pursuing them further. The response to the present SOAR will provide the future directions that our work will take.

We have completed our current voyage through statistical analysis. We hope the reader has been able to acquire a better appreciation and understanding of the statistical thinking behind the mechanics of the statistical procedures discussed. We also hope that our readers have develop more interest for further study of statistics, a fascinating subject that allows one to better deal with and extract information from, ever-present data.