THEORIA ET HISTORIA SCIENTIARUM, VOL. VI, N° 2
Ed. Nicolas Copernicus University 2002

Sallie Keller-McNulty, Mark S. McNulty

Show Me the Data: Statistical Representation

Abstract

Statistical representation is the science and art of using data to describe the world around us. Statistical representation is based on the fundamental concept that data consists of structure plus noise. The challenge facing the statistician is to use the noisy data to learn about the underlying structure. This framework accommodates the analysis of data generated by almost all other scientific disciplines. There are numerous ways of constructing statistical representations. The methods discussed here include tables, graphs, and models. The proper representation depends on the nature of the data and the particular issues being addressed. A combination of methods is often appropriate.

1. Introduction

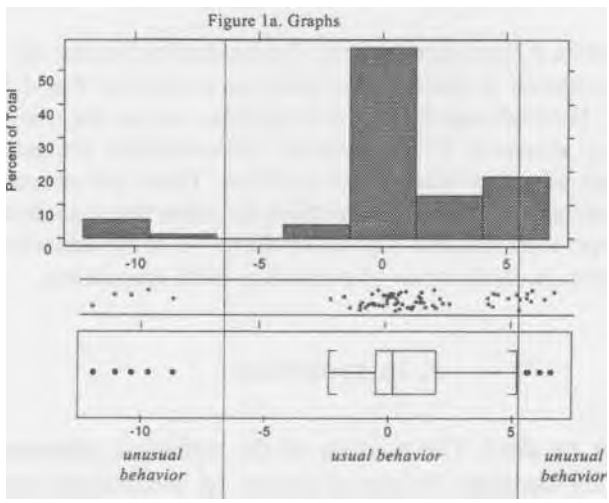
Statistics is driven by *data*. The mission of the statistical sciences is to serve science and society through the development of techniques for collecting, summarizing, and making inferences from data. It is an explosive time for statistics, with no shortage of novel data in need of analysis. Statistics has become the quintessential interdisciplinary science because scientists and engineers from other disciplines collect the data and formulate the problems statisticians seek to solve. Statisticians reach out, intentionally and enthusiastically, to all areas of science and engineering in the pursuit of interesting and important problems to solve.

The focus of this article is on statistical representation, which is about using data to describe the world around us. Data are generated in many ways (surveys, interviews, sensors, ...), come in a multitude of formats (text, numbers, sounds, images, ...), and can have a variety of dimensions (spatial, temporal, logical, ...). Despite the rich and varied nature of data, there is a fundamental concept

underlying the statistical use of data to represent the world. This concept is that the process that generates the data consists of a regular, predictable component (structure) and a random, unpredictable component (noise):

$$\text{data} = \underset{\substack{\uparrow \\ \textit{usual behavior}}}{\text{structure}} + \underset{\substack{\uparrow \\ \textit{unusual behavior}}}{\text{noise}}$$

This simple framework applies to all scientific disciplines. The structure is what we strive to understand. The noise is a nuisance that we must deal with. The statistician's task is to wipe the noise from the data so that what remains is a clear vision of the structure. This rest of this article will focus on using data to construct a statistical representation of the structure using simple graphs and models (Figure 1 represents the discussion that will follow).



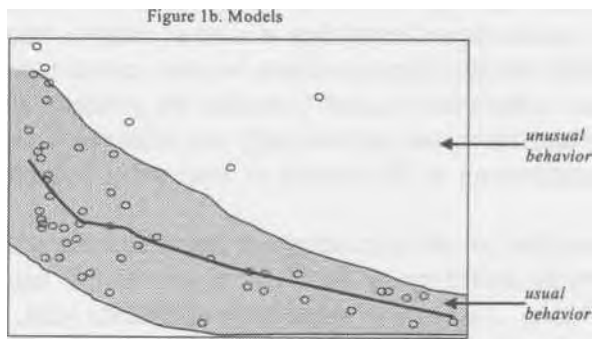


Figure 1. Statistical representations

The particular slice of the real world that is being studied will be referred to as a *system.* The term "system" is used in a very general way in this discussion, and just about any piece of the world one wishes to specify or imagine can be considered a system. An important use of statistical representation is to provide a concise description of a system. For example, the descriptive statistical graphics in Figure la illustrate the range of possible data values and some idea about which values are likely to occur and which values are unlikely. A more complex representation is the scatter plot in Figure lb, in which a statistical model has been used to describe the relationship between two elements of the system. Here statistical methods were used to find an algebraic definition of the line that "best fits" the data and that summarizes the relationship between the two plotted variables. Both of these examples make clear the importance of data in statistical representation. Data are the fundamental source of information the statistician has about the system being studied.

2. Data

Before discussing how data are used, it is useful to describe some of the different types of data that may occur. One important distinction is whether data are quantitative or qualitative. Quantitative data arise when measurements are taken on quantities. For example, the number of visits to a web site, voltage through a wire, and income of a person would be quantitative data. Qualitative data arise when measurements indicate whether some entity possesses a certain quality. The domain of the web site (that is, .com or .edu), whether the wire is copper or aluminum, and a person's gender would be qualitative data. Qualitative data must be handled differently than quantitative data. One obvious reason for the distinction is that algebraic concepts (addition, subtraction, greater-than, and so forth) cannot be meaningfully applied to qualitative data. A given study will often make use of both quantitative and qualitative data. Using both types of data simultaneously creates interesting challenges.

Another important data characteristic is whether the data are observational or experimental. Observational data are generated by observing the world as it progresses, without any external influence by the scientist. Experimental data are generated in a controlled setting in which certain factors are set at predetermined levels and others are allowed to vary. The advantage of experimental data is that the noise component of the data is controlled, and it is easier to identify the factors that determine the structure. The disadvantage of experimental data is that they are typically more costly to generate than observational data. In many cases experimental data are impossible to collect altogether. This is especially relevant for one of the most important areas of

study, that of the human condition. Constructing experimental settings in which people behave in a natural way is very difficult. Consequently, most societal data are observational. Given the unpredictable manner in which most people behave, the noise component of such data is often very large relative to the structural component. It is not unusual for 90% or more of the data taken on individuals to be unexplainable noise. Sifting the grains of structure from such data is the kind of challenge that makes statistics rewarding.

Quantitative versus qualitative, observational versus experimental, data source, data format, and data dimension are just a few of the different characteristics data may possess. With our increasing ability to gather large amounts of data, an increasing problem is keeping track of what, exactly, the data are. An important advance in computer science is the development of databases that keep track of such information. In addition to the actual data, these databases contain schema and metadata. Schema are the logical structure of the database. Traditionally, metadata has been thought to describe the format or layout of the data. A new form of metadata has recently emerged, *statistical* metadata, which are descriptive information or documentation (for example, the sampling plan or imputation method) about the data that greatly facilitates using and sharing the data. The metadata can now be as voluminous as the original data. This has led to the creation of metadata databases. Research is ongoing to develop statistical representations of metadata and methods to mine metadata (Wegman 1999).

Given a set of data, the statistician is faced with the task of saying something about the structure of the system that generated the data. Scrutinizing a list of data records is not very productive. Making sense of a list of even 30 data points pushes most of our capabilities to the limit, and most data sets are far larger than that. Somehow these long lists must be reduced to a manageable set of values. The intuitive solution to this problem is to aggregate the data. When the data are quantitative the natural inclination is to compute averages of the data and to examine those values. This turns out to be precisely the correct strategy, given the structure plus noise model of the data. The reason is that the averaging process minimizes the noise component of the data and magnifies the structural component. This result is so powerful that the vast majority of all statistical computations are based upon averages. In order to measure different aspects of the underlying structure the data may be transformed in various ways and the weighting may not be simple, but in the end an average is indeed taken.

The averaging operation cannot be applied directly to qualitative data; averaging the values MALE and FEMALE gives a nonsensical result. With qualitative data, a natural way to aggregate and summarize the data is to create a table based upon the different possible classifications, and to compute the fraction of the data that fall in each classification. This computation is a type of

average and has the same desirable property of magnifying structure. The structure in this case is the tendency of individuals to fall in one table cell or another. While a table constructed for a single variable can be quite useful, it provides no information about a key element of structure: relationships. The description of the relationships within the system is often a primary objective of statistical representation. Tables can be used to fulfill that objective if they are based on the cross-classification of two variables. Of course, tables can also be constructed using quantitative variables by defining value ranges to be the classification categories.

3. Graphs

Graphs have historically been a powerful mode of statistical representation. In the previous section, we noted that the ability of humans to interpret lists of data is very limited. However, our ability to process graphical displays is outstanding. Indeed, nothing can outperform human visual capabilities for pattern (structure) recognition. Graphical displays help to isolate patterns and features that are worthy of further study, uncover unexpected behavior, or lend support for expected behavior of the system under study. Graphical displays also play a role in helping the analyst determine how rich a model the data can support.

Some of the simplest displays are frequently the most useful. The histogram, one-dimensional scatter plot, and boxplot (see Chambers et al. 1983 for an excellent description of these and other graphical displays) in Figure la make it immediately obvious that the data have a large degree of skew and that statistical methods requiring an assumption of underlying normality (Gaussian distribution) are likely to fail. These displays can be easily used to compare subsets of data. Figure 2 demonstrates this concept with AIDS incidence data for 38 U.S. metropolitan statistical areas (MSA) from 1983 to 1990.

Two-dimensional scatter plots are arguably the most widely used graphic by all of science. The likely reason is that they represent system relationships clearly and directly, and relationships are usually the focus of attention. Variations on the scatter plot theme can be used to identify unexpected structure. For example, Figure 3 is a scatter plot for five of the MS As in Figure 2. Each MSA has been plotted with a different symbol and connected by the solid lines. Comparing Figures 2 and 3, it is apparent that the most extreme incidence counts in each year come from the same MSA.

Low-dimensional tables of data can also be converted to useful graphical displays. For example, Figure 4a displays pairwise comparisons between groups, highlighting the control group comparison. The box heights represent the pooled standard error of the group mean. Figure 4b displays the mean response and

standard error for a 3x3-treatment design. Again the height of the bars represent the standard errors. Figure 4b clearly represents the changing trends in both levels of the treatment structure. The information in these examples can simply not be conveyed as clearly using tables of means and p-values.

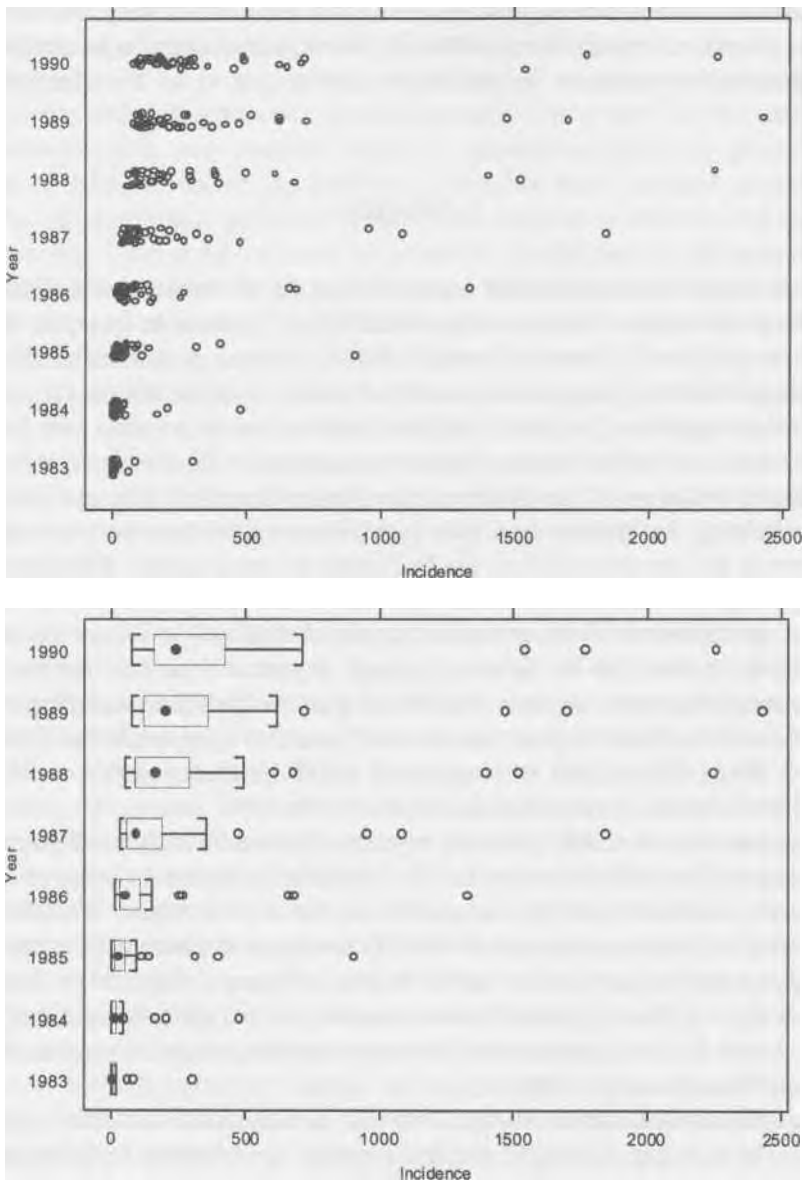


Figure 2. One-d scatter plots and box plots of AIDS incidence from 1983 to 1990

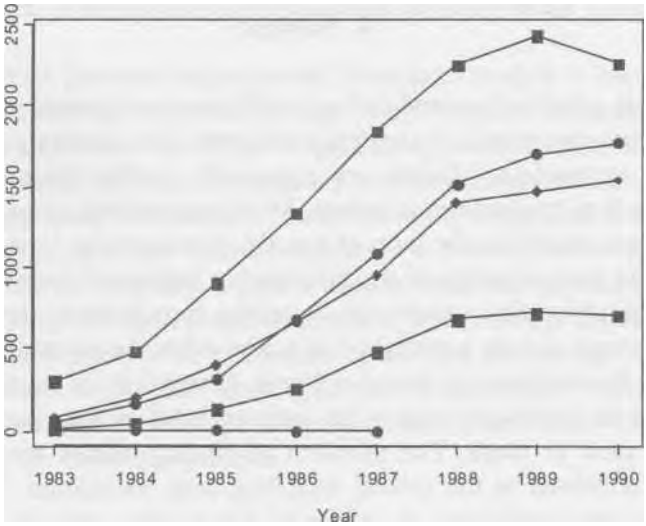
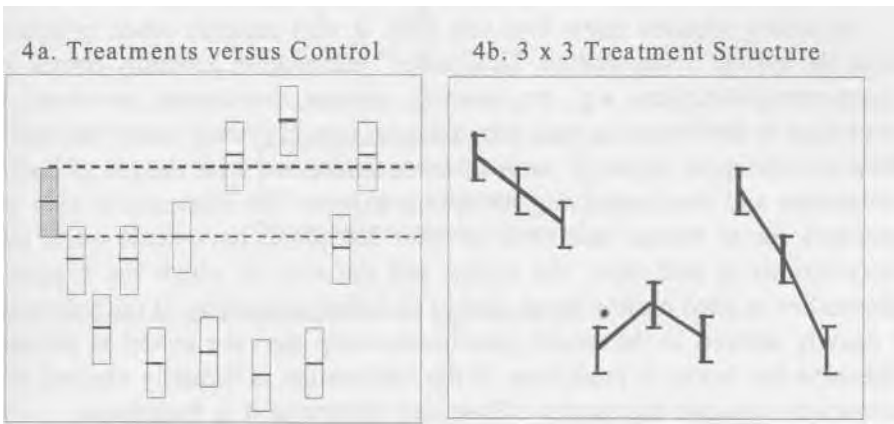


Figure 3. Time plot of AIDS incidence for five MSAs from 1983 to 1990



.- .gure 4. Graphical display of tabular data

Three and higher dimensional statistical graphics and dynamic graphics, including stereo displays and virtual reality, are very active areas of research (Wegman 1998). In statistical sciences, this research frequently has a cognitive psychology component (Cleveland 1985) which seeks to assess the usefulness if the
displays vis-a-vis human cognition.

4. Models

Statistical graphics go hand-in-hand with the development and verification of algebraic representations of a system's structure. Such algebraic representations are known as models. Models are especially useful for quantifying the relationships that exist within a system. Modeling consists of two major steps. The first step is specifying the form of a model. For example, it might be assumed that a straight line summarizes the relationship (structure) between salary and education. Graphics play a major role in model specification. One would never (or, should never) specify a straight-line model when the scatter plot of the data showed that the relationship was curvilinear. Given that the model is a straight line, what is its slope and what is its intercept? All models contain unknown parameters such as these. The problem of finding values for the unknown parameters is solved in the second modeling step, estimation. Estimating, or fitting, the model entails using the observed data to infer values for the unknown parameters. While the details can become messy, the basic idea has an appealing visual interpretation. The chosen parameter values give a model that, when plotted with the data, gives the "best fit", as illustrated in Figure lb.

Modeling requires more than just data. It also requires other information about the system being studied. This "other" information generally comes from a supporting discipline, e.g., engineering, physics, economics, sociology, that specializes in understanding how that particular type of system works. Statisticians often become quite expert in some of these fields, and wear the hat of both the statistician and the supporting discipline expert. The converse is also true; scientists whose formal training is in other disciplines have made many major contributions to statistics. The extent and the way in which the supporting information is used creates broad classes of statistical models. If the information is heavily utilized in the model specification step then the model is structural, otherwise the model is predictive. If the information is formally utilized in the estimation step then the model is Bayesian, otherwise it is frequentist.

The naming convention for structural and predictive models is a little misleading. Both types of models are used for prediction. Recall that if the information from the supporting discipline is incorporated into the model specification then the model is structural. If little supporting information is used during the specification phase then the model is predictive. The idea is that a structural model actually represents the inner workings of the system as understood by the supporting discipline. With a predictive model, whether the model actually mimics the real world is not an issue. The only concern is how well the model predicts. Of course, because structural models use more

information one would hope that they would predict better than predicative models. However, predictive models sometimes perform better because of their simplicity.

The distinction between Bayesian and frequentist models is based upon the way in which the information from the supporting discipline is incorporated into the model estimation step. Under the frequentist approach estimation is entirely based upon the data; the final estimates are functions solely of the data. Under the Bayesian approach, estimation is based formally upon both knowledge of the system and the data; the final estimates are a blend of what the Bayesian believes and what the data say. Figure 5 depicts these two approaches, both of which have strong advocates and opponents. Frequentists argue that their methodology is less likely to be influenced by the bias and predisposition of the analyst. Bayesians argue that their approach utilizes more information in a legitimate manner and therefore produces better model estimates. There is likely an element of truth to both arguments.

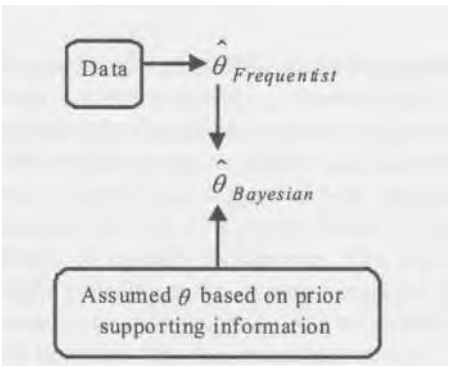


Figure 5. Bayesian versus frequentist parameter (g) estimation

6. Conclusions

Statistical representation has a solid, and sometimes very abstract, theoretical foundation. However, the methods used generally have a strong intuitive appeal. This is especially true of graphical displays, one of the most powerful tools of statistical representation. If a graph does not make intuitive sense it is not a good graph. While the details of statistical modeling can become quite complex, the relationships and structure that the model seeks to reveal and quantify are generally of obvious interest to even a casual observer. The methods of statistical

representation are tools developed for use by other scientists. Without a strong interaction with the rest of science, statistics becomes somewhat of an empty exercise.

References

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey (1983). *Graphical Methods For Data Analysis*, Duxbury Press: Boston.

Cleveland, W. S. (1985). *The Elements of Graphing Data*, Duxbury Press: Boston. Wegman, E. J. (1999). "Visions: The Evolution of Statistics," *Research in Official Statistics* 1, 7-19.

Wegman, E. J., Luo, Q., and Chen, J. X., (1998). "Immersive methods for exploratory analysis," *Computing Science and Statistics* 29(1), 206-214.