

STATISTICAL GRAPHICS: PHILOSOPHY TO THE RESCUE!

Elena N. Naumova and Eileen A. O'Neil

Tufts University School of Medicine, Boston, Massachusetts 02111

KEY WORDS: Statistical Graphics, Philosophy, Ethical Guidelines, Discrete Distributions, Polyclonal Immune Response

ABSTRACT

The graphical display of data is a burgeoning field. Graphic visualization gives rise to new statistical challenges and a host of intriguing questions related to several areas of philosophy: semiology, epistemology, ontology, and ethics. In this venue, we will explore the role of language, the significance of definitions, and the ethical challenges in constructing a graphical display.

By its nature, statistics relies upon numbers; philosophy's tool is words. Although the field of statistics is intricately connected with philosophy, both historically and conceptually, "doing philosophy" can be an exercise in frustration for non-philosophers. However, engaging in philosophical questioning can be enormously rewarding; ironically, frustration can lead to clarity, enabling statisticians to proceed with confidence.

In this paper, the interrelationship of philosophical questions and a graphical display is demonstrated by an example from immunology, which begins with stating a hypothesis, continues with data collection, and moves to a graphical presentation. Based on our philosophical analysis we conclude: 1) a graph must rest on and adhere to proper definitions and a set of operational rules and research methodologies, 2) translating a display into a language can serve as a standard to determine if a graphical presentation is overstating results, or has overemphasized a point, or is suggesting more than the data support, 3) if a graphical display is acting as a language, a tool for communication, to avoid miscommunication one should create a set of rules, operations, and terms, essential for a language.

A simple rule of graph construction is that a graph without a well-understood statistical context or logical path, without good visual properties, and without cognizance of the audience is hardly worth drawing.

1. INTRODUCTION

A graphical display is a product of data visualization, a process of forming mental images of the data, reflecting the data's main features and the

relationship among data elements. As a product of cognitive processes, a graphical display done correctly will contain knowledge from the research and convey meaning to others. Therefore, a graphical presentation is a very powerful communication tool. The widespread availability of "do-it-yourself" computerized gizmos for data visualization stimulates an enormous production of graphs, charts, plots, maps in both public media and research literature in practically all fields of science with no guarantees of quality.

The variety of graphs, charts, plots, maps, and diagrams is unlimited, but they all share common features and are limited by a two-dimensional reflective medium. The shared features, the similarities in a variety of graphical data images, form the basis of preliminary classifications of graphical displays. These similarities can be of two kinds: 1) the similarities of graphical displays in their relation to a viewer, in our visual perception or their visual appearance, and 2) the similarities of graphical displays in their relation to each other.

Applied statisticians continue to face a broad range of traditional challenges related to: first, the dependence of statistical inferences on the soundness of research hypotheses and working definitions in a field of application, and second, the quality of study design and collected data. Exciting opportunities to make and deliver attractive graphical displays quickly and easily give rise to the new challenges and to a host of intriguing questions: How is a reliable graph produced? How is a graphical display evaluated? What does a graphical presentation say to the public? What does the intended audience read into a graphical presentation? What are the conditions under which an inference from a graphical presentation can be drawn?

In a search for an insight into developing a systematic way to address challenges related to data visualization and to avoid the traps that may compromise professional contributions of a statistician to the field of science, we decided to use philosophical analysis and work with several fields of philosophy: epistemology, ontology, semiology, and ethics, which deal with different subject matter and therefore consider different questions.

In this paper, we raise and discuss a range of philosophical questions that arise from a hypothetical statistician's "work-in-progress" example presented in delineated steps. The example that runs throughout the paper was borrowed from the study of human immune response to influenza virus (Naumov *et al.*, 1998). It

starts with setting a hypothesis, moves to data collection and analysis procedures, and ends with a simple distributional representation. We illustrate our philosophical reasoning by assuming that a “great idea” along with “great data” can lead to a “great display”.

We conclude our travel to the worlds of graphical display and philosophy with practical recommendations for drawing, dreaming and wonder.

2. PHILOSOPHY TO THE RESCUE!

The disciplines of statistics and philosophy seem to be drastically different in scope, nature, and history. Statistics relies on numbers whereas words are the main tool of philosophy. Statisticians, especially those who work with scientists, are closely tied to the empirical world, while philosophers deal with the abstract.

Philosophy is as ancient as the human ability to wonder, reflect, question, and investigate with the mind, as revealed by its Greek roots *philo-* love of + *sophia* wisdom. Its questions cannot be addressed by empirical verification. In fact, philosophy uses only rational discourse to achieve its end. Yet language is equally important to statistics as it is to philosophy. Despite the frustration that philosophy can bring to the uninitiated, the intersection of philosophical analysis with statistical practice can be very useful and even entertaining.

Philosophy has several areas that deal with different subject matter and therefore consider different questions. Epistemology is the area of philosophy dedicated to the nature and limits of knowledge. For example, can we know nature as it is, or are we constrained by our inherent human limitations? Are there eternal truths, and under what conditions can we understand them? Semiology is the study of the nature of signs and symbols. Of concern to a semiologist would be “to what does the sign ‘X’ refer?” - not in terms of its actual mathematical value, but the nature of the mathematical value. Ontology is the study of “being” and encompasses a question such as “does a number belong to the corporeal or incorporeal world?” or “what do we mean when we assert that something is ‘real’?” Ethics delves into values, entertaining questions such as “what is the moral life?” and “how should statisticians conduct themselves in an ethical way in graphical presentation of data?”

The nature of these questions is an indication of the reasoning in philosophy that differs from the way one approaches statistics, immunology, painting, or common sense.

3. PHILOSOPHICAL MUSING ON POLYCLONALITY OF THE IMMUNE RESPONSE

To raise and discuss a range of philosophical questions that can arise from an applied statistician's “work-in-progress”, we have borrowed an example from the recent immunological discovery of polyclonal response to influenza peptide. To illustrate our philosophical reasoning we are presenting this example in three delineated steps: from “great idea” through “great data” to “great display”. We supply each step with philosophical commentary and then translate those comments into statistical remarks at the end of the paper.

Step 1. The Research Hypothesis: “the great idea”

The first step reflects the process of formulating research hypotheses and working definitions on which statisticians will rely in their work. Below we describe how an immunologist states the research hypothesis about an anticipated mechanism of the polyclonal immune response.

*Example: The manner in which naïve T cells give rise to the immune memory repertoire, how the repertoire is sustained, and how it changes in time, represents an example of a complex system. Assuming that high avidity T cells will be selectively expanded following successive challenges with antigenic peptide presented by the major compatibility complex (MHC) molecules, it was predicted that a memory repertoire would contain multiple copies of T cells expressing limited diversity of T cell receptors (TCR). **Figure 1** below reflects an example of the crystal structure of the TCR-peptide-MHC complex.*

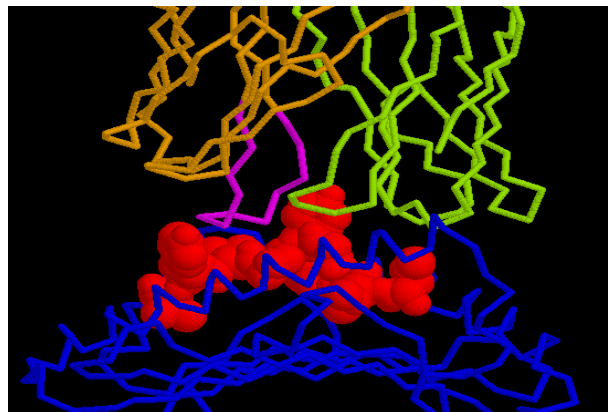


Figure 1. The crystal structure of the TCR-peptide-MHC complex: TCR (shown in brown, green and magenta) interacts with the antigenic peptide (red) loaded into the MHC class I molecule (blue).

Three commentaries related to the nature of the Hypothesis and soundness of definitions are provided below.

Commentary 1. The Nature of the Hypothesis

A sound hypothesis is the lynchpin of good science. Philosophers of science take various views of the way in which science actually progresses. Some argue that science germinates from an abstract research idea, followed by data collection and analysis to support or refute the hypothesis. In contrast, the more empirically inclined philosophers propose that the guiding scientific hypothesis develops from observation of data. The dispute on the relationship of data and hypothesis in this sense dates back to the ancient Greek philosophers Plato and Aristotle. In Plato's view, the abstract "Ideas" or "Forms" - for example, "Love" as a concept, as opposed to individual instances of love - are the most "real" since the Forms endure through time without being subject to the changes that occur as a result of their material nature. Aristotle, to the contrary, argues that the development of a hypothesis necessarily begins with the sensory data used to form the definitions (Aristotle, Posterior Analytics 71b20-72b2).

Commentary 2. Soundness of the Hypothesis

The soundness of the scientific premise is required to arrive at the truth or falsity of the premises. For example, a scientist could begin with the hypothesis "pigs with wings are blue," and conduct research that assumes that data will substantiate the claim. However, the results will be unsound since we know from experience that pigs lack wings. Soundness, then, is a critical component of the research hypothesis since there is a danger that the researcher could interpret the data in accordance with an unsound premise. For example, it is possible that a researcher has an incorrect notion of the T cell receptors. Without an accurate scientific concept, the research is doomed to fail. How, then, is soundness achieved at the hypothesis step? Virtually all philosophers of science argue for the necessity of solid definitions of the terms used in the hypothesis. In the Aristotelean view, echoed by many contemporary philosophers of science including Bernard J.F. Lonergan, a proper definition of a thing develops from the process of abstracting the essential attributes of sensory data (Lonergan, 1957). We can see active nature of this process in Lonergan's characterization: "...for the image as image can be only verified by the occurrence of corresponding sensations..." What is produced by induction is the "universal," the "essence" of things or processes which includes in its purest form the "whatness," or the "why" of it (Aristotle, Physics 194b20-195a4). For example,

the "immune response" is a "universal" developed from observation of the particulars with their variations. Leaving behind these individual differences, the essential attributes of "immune response" form its definition, a "universal" applicable to humans in general, which allows research to be conducted.

Commentary 3. Soundness of the Definitions

The excellence of a scientific hypothesis depends upon the excellence of the definitions. If the terms are improperly defined, then the entire endeavor - including the statistical analysis - falters. If a definition of "clonotype" is not accurate, the more complex statement containing the term will not produce knowledge. As discussed above, in the strictest sense, the proper definition will include an understanding of the "whatness," or the reasoned cause.

Step 2. An Accurate Methodology: "the great data"

The researcher must select and apply an appropriate and accurate methodology for obtaining data. This step determines the data quality that is essential for a statistician in forming statistical assumptions and selecting procedures for analysis. An exquisitely executed statistical analysis performed in this step is necessary in order to generate valid research conclusions.

***Example:** To analyze the memory T cell repertoire, two cultures ("A" and "B") were generated from the peripheral blood mononuclear cells of an individual with the strong immune response to the influenza A virus peptide M1₅₈₋₆₆. The unit of measure of the T cell repertoire is a clonotype, represented by the unique DNA sequence that encodes the complementary determining region 3 (CDR3) of TCR β -chain and is quintessential for peptide recognition, or avidity. In Figure 1, the CDR3 regions are the segments of the TCR, which are in the closest contact with the antigenic peptide (shown in green and magenta). The clonotypes were identified by the DNA polymerase chain reaction of the TCR β -chain CDR3, subcloning and sequencing. The scheme of recording the presence of a clonotype is shown in **Figure 2**. Analysis of both cultures identified 294 bacterial colonies containing TCR inserts and these were accounted for by 95 unique clonotypes.*

Commentary 4. Validity

We know from the principles of formal logic that the solid research requires both validity of methods and soundness in the premises as we discussed above. Validity is achieved by proper inferences drawn from the premises.

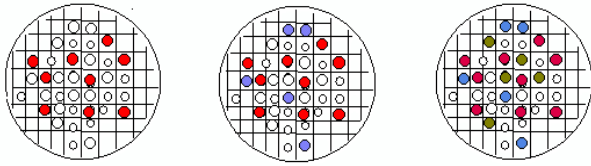


Figure 2. The sequential scheme of recording the presence of a clonotype is shown in three stages of identification. The colored dots (red, blue and green) reflect the presence of 20 colonies with the inserts of three unique clonotypes.

Commentary 5. Inferences and Causality

A critical issue in science, stemming back to ancient Greek thinking, concerns the conditions under which data can be said to be causally connected. In contrast with some branches of science, like epidemiology, most philosophers of science would argue that knowledge is composed of both statistical correlations and a classical explanation that contains the “why” of the connection between events. Once understood, the “why” becomes incorporated into the definition of the thing, event or idea.

Step 3. Data Visualization: “the great display”

Visualization is the process of constructing an informative view of the data appropriately grounded in a statistical context. Visualization is critical in understanding and describing subject-specific properties, structures, and relations and is also critical in communicating the derived information, query, or statement to both professional and lay audiences.

Example: Each unique clonotype was represented by a number of colonies. When the relative frequencies of colonies for culture “A” were plotted in descending order (Figure 3), it revealed that a few clonotypes contained multiple copies and the majority of clonotypes were presented by only a single copy, which formed a long tail. To further describe properties of the clonotype distribution we assigned each clonotype a rank based on the absolute counts of copies (rank 1 consisted of clonotypes observed as single copies, rank 2 those observed twice, etc). When the relative frequencies for each rank were plotted in increasing rank order, it revealed a power-law-like rank-frequency relationship (Figure 4). The graph of relation between the log-transformed rank and the log-transformed frequency is shown in the inset of Figure 4. To make the scale on the axis easy to compare, the base of log-

transformation was taken as the maximum observable value of the rank and absolute counts. The fitting of the power law curve was performed satisfactory for the first part of the curve; but the tail fitting required a special treatment.

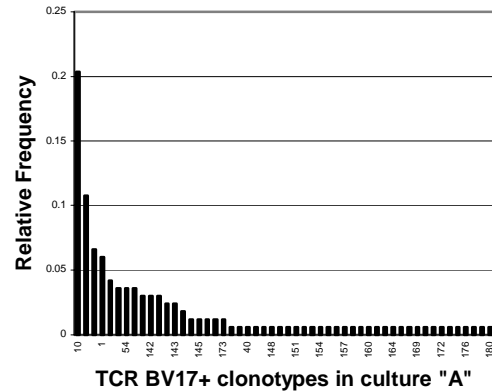


Figure 3. The relative frequencies of colonies for culture “A” plotted in descending order.

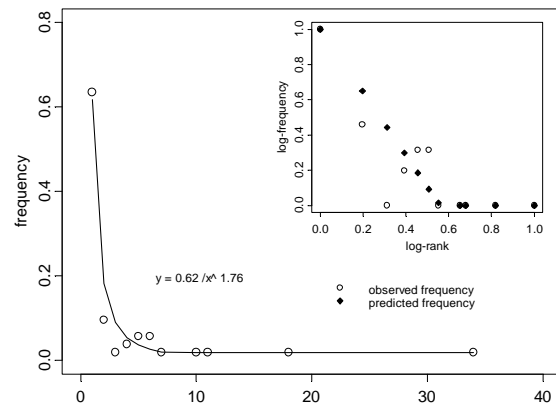


Figure 4. Frequency of frequencies distribution for culture “A”. In the inset, the relation between log-rank and log-frequency is shown. The predicted values were obtained by fitting the linear regression model, parameters of which are also shown in the graph.

Although the power-law-like shape of the observed clonotype distribution is predetermined by the fact of sorting the clonotype relative frequencies in the descending order and, therefore is expected, the relations among the different fractions of this distribution are not apparent. By allowing the abstraction and grouping by rank, the observed power-law-like shape in the rank-frequency would be of

research interest. This technique of converting data to a frequency of frequencies distribution is well described in statistical literature related to the analysis of species abundance (Good, 1953; Hill, 1974).

Commentary 7. Visual Perceptions

In contrast with a tabular presentation of data, a viewer initially grasps a graphical display in its totality, with all its attributes. One philosophical perspective on this issue is given by Bernard J.F. Lonergan: "...the flow of sensations as completed by memories and prolonged by imaginative acts of anticipation becomes the flow of perception..." Sensible consequences of the visual perceptions initiated by selected color, size, or style of presentation can give knowledge in its highest form or lead astray.

Commentary 8. Ethics

To discuss ethics with regard to a graphical display requires recognition of the intellectual virtue of knowledge. Knowledge is the aim toward which all decisions are made, including the use of carefully formulated definitions as well as the selection of the attributes of the graphical presentation that may, at first, seem inconsequential.

4. THE TRAPS AND CHALLENGES OF GRAPHICAL DISPLAYS

Remark 1. Soundness of Hypothesis and Definitions

In applied statistics it is assumed that scientists in the appropriate fields of applications properly define the things, events and processes. The statistician, then, can be said to be dealing with the definitions as final products. Rather than testing the definitions themselves, a statistician will be testing the hypothesis that contains the definitions. However, as has been shown, research is only as sound as the definitions of terms in the hypothesis. On occasion, a statistician may recognize from a variety of sources that the terms in the hypothesis were not properly defined. Under those conditions, the statistician should make inquiries in order to continue with the research, for this initial flaw jeopardizes the entire enterprise.

A reliable graph can be made only if the process of abstraction from empirical observations has reached a proper definition. Furthermore, graphical presentations based on research using incorrect or haphazard definitions can confuse or mislead the viewer, violating a basic ethical premise. The statistician should consider how the research presentation may give the wrong impression and at the extreme, cause harm.

Remark 2. Validity and Inferential Reasoning

The accuracy of the research results depends not only on proper definitions of terms - those that contain the "whatness" - but also on the selection of the appropriate methods to obtain data, appropriate statistical methods to apply to the data, and appropriate graphical tools to demonstrate the results. In applied statistics it is assumed that the researcher would apply appropriate methods to obtain data, so the statistician is again dealing with data as final products. Needless to say, reliable statistical analysis and graphs can be made only if data are correct. It is well known that data visualization, as exploratory data analysis, can be exceptionally helpful in identifying implausible or incorrect data.

On other hand, knowing the "what, how and why" about collected data, which are the subjects for analysis and visualization, is critical for constructing efficient and reliable graphs. Graphical data images can be created by using a wide variety of graphical presentation tools such as graphs, charts, plots, maps, and diagrams. In selecting a data visualization technique the following should be considered: appropriateness, accuracy and visual perception of the graph.

Remark 3. Role of Language and Completeness

The role of language in a graphic display can be dual: a graphical display can have some properties of a language itself and at the same time it can serve as shorthand for language. Furthermore, a graphical presentation and its verbal explanation should be co-extensive. Therefore, a presentation standing alone should conform to the explanation given by the researcher or statistician in words.

If the researcher needs to qualify the visual display, that is, must explain to the audience in words what the display does not say, or if the graphical presentation says more than the research, then the graphical display may be misleading. A proper graph, if translated into a language, should clearly reflect the data and research results without losing details or overemphasizing a point.

Remark 4. Visual Perception and Plausibility

Symbols play a special, dual role in a graphical presentation. A graph can be viewed as a symbolic representation of a thing, event, or process, and on the other hand, symbols are essential attributes of a graph. In both cases the function of the symbol is to supply the relevant and reliable image. As Lonergan pointed out, "...there is no doubt that, though symbols are chosen by convention, still some choices are highly fruitful while

others are not." In the process of symbol selection for a graph one should consider that symbols offer clues, hints, suggestions. The usage of symbols might become customary and dictate inference. An unexplained change of a pre-selected symbolic pattern over the course of graphical presentation might lead to confusion and give the wrong impression. If in the process of graph construction, the impact of selected symbols, color, scale, size, and style are not taken into account, a graphic display might not only lose its efficiency but even mislead the viewer.

Remark 5. Ethical Challenges

Ethics within our context involves the myriad decisions that a statistician faces. A statistician who thinks that the terms are not properly defined, or data quality is under suspect, must make a decision to either proceed with the analysis, or – often a more difficult path – set limits on what can be done with the data.

Many of the provisions in the statistical practice code speak to other specific kinds of obligations, which appear in the ETHICAL GUIDELINES FOR STATISTICAL PRACTICE established by the American Statistical Association (www.amstat.org).

A different kind of ethical challenge is assuring that the graphical presentation reflects the results and does not tell a story that is not true.

5. SUMMARY

As the result of our philosophical analysis we think that 1) a graphical display must rest on and adhere to proper definitions and a set of operational rules and research methodologies, 2) translating a graphical display into a language can serve as a standard to determine if a presentation overstates the results, overemphasizes a point, or suggests that more is known than the research supports, and 3) if a graphical display is acting as a language, and therefore as a tool for communication, to avoid miscommunication one should create a set of rules, operations, and terms, essential for a language.

Although there are some rules, which are intuitive and not operational, they could provide some insight into how to create a well-constructed graphical display: a graphical display should contain a well-understood statistical context or logical path, for which one is able to give a verbal description; should help to explain data or concepts by taking advantage of visual perceptions; and should force one to note the unexpected, motivate questions, and clarify statements, results or concepts.

Although impressive attempts to develop the guidelines for constructing good graphs have been made (Wilkinson, 1999; Harris, 1996), there is a need to explore the symbolic nature of graphical displays, where a graph's key elements and attributes comprise its language and style. ♦♦♦

ACKNOWLEDGMENTS

Authors wish to thank Drs. Yuri Naumov, Kevin Hogan and Jack Gorski for providing original data on polyclonal immune response and their thought-provoking questions, suggestions and comments as well as Pamela Foster for editorial assistance.

REFERENCES

ARISTOTLE, PHYSICS (Translated by Hippocrates G. Apostle. 1969)

ARISTOTLE, POSTERIOR ANALYTICS (Translated by Hippocrates G. Apostle. 1981)

GOOD IJ. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*. 40: 237-264.

HARRIS RL. (1996). Information graphics: a comprehensive illustrated reference. Management Graphics. Atlanta, Georgia.

HILL BM. (1974) The rank-frequency form of Zipf's law. *JASA*. 69:1017-1026.

LONERGAN B. (1957) *Insight: A Study in Human Understanding*. Philosophical Library, New York.

NAUMOV YN, HOGAN KT, NAUMOVA EN, PAGEL JT, GORSKI J. (1998) A Class I MHC Restricted Recall Response to and a Viral Peptide is Highly Polyclonal Despite Stringent CDR3 Selection: Implications for Establishing Memory T-cell Repertoires in "Real-World" Conditions. *The Journal of Immunology*. 160, 6: 2842-53.

WILKINSON L. (1999) *The Grammar of Graphics*. Springer-Verlag. New York.