Visualization of Statistical Measures of Uncertainty

Kristin Potter*

School of Computing, University of Utah

Abstract

One of the important problems facing visualization is the lack of uncertainty information within the visualization context [2]. This work investigates the problem starting from a graphical data analysis standpoint. By using descriptive statistics to investigate uncertainty, the goal of this work is to clearly present the quantifiable aspects of uncertainty, rather than the *unknowns* of uncertainty.

Keywords: Uncertainty, Descriptive Statistics, The Box Plot

1 Introduction

While visualization effectively presents large amounts of information in a comprehensible manner, most visualizations typically lack indications of *uncertainty*. Uncertainty is a term used to express descriptive characteristics of data such as deviation, error and level of confidence. These measures of data quality are often displayed as graphs and charts alongside visualizations, but not often included within visualizations themselves. This additional information is crucial in the comprehension of information and decision making, the absence of which can lead to misrepresentations and false conclusions.

2 Problem Statement

Most visualization techniques that incorporate uncertainty information treat uncertainty like an unknown or fuzzy quantity; a survey of such techniques can be found in [3]. These methods employ the syntax of the word uncertainty to create the interpretation of *uncertainty* or *unknown* to indicate areas in a visualization with less confidence, greater error, or high variation. Blurring or fuzzing a visualization, while accurately expressing the lowered confidence one should have in that data, does not lead to a more informative decision making tool, but instead obfuscates the information that lead to the measure of uncertainty. Such a solution to the problem of adding qualitative information to visualization does not elucidate on the quantitative measures leading to the uncertain classification, and thus is missing some important information.

The goal of this work is to identify and visualize the measures typically thrown under the umbrella of uncertainty. Uncertainty, as the scientific visualization field titles it refers to quality of a measured value and can include measures of confidence, error, and deviation. Statistically, uncertainty is harder to identify, since there are many measures that can add to the qualification of data. Using measures that are statistically meaningful to express the uncertainty in a data set exposes insights to the data that may not have previously been obvious. Adding such quantities to visualizations will improve the effectiveness of visualizations by providing a more complete description of the data, and create better tools for decision making and analysis.

3 Approach

Exploratory data analysis, as coined by John Tukey, uses graphical techniques to summarize and convey interesting characteristics of a data set to facilitate not only an understanding of the given data, but for further investigation and hypothesis testing. These tested graphical methods, such as the Box Plot, Histogram, and Scatter Plot, are easily identifiable signatures of a data distribution, and their simplicity allows for quick recognition of important features, comparison of data sets, and can be substituted for the actual display of data, specifically when data sets are too large to efficiently plot.

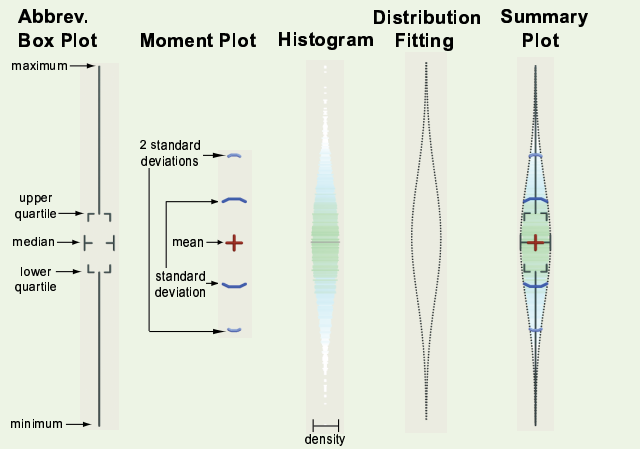

Figure 1: The Summary Plot on a Gaussian distribution.

This work takes inspiration from the visual devices used in exploratory data analysis and extends their application to uncertainty visualization. The statistical measures often used to describe uncertainty are similar to measures conveyed in graphical devices such as the Histogram and Box Plot. The first step of this research investigated the creation of the *Summary Plot* which combines the Box Plot, Histogram, a plot of the central moments (i.e. mean, standard deviation), and distribution fitting. The success of this work has lead to the hypothesis that similar visual techniques can be used in higher dimensional data, by using visual metaphors to indicate a variety of statistical measures as an overlay to traditional visualization techniques.

One problem faced when supplying additional information in a visualization is visual clutter and information overload. This can quickly become a problem in this approach because of the variety of available visual information, and the desire to allow traditional visualization techniques to be incorporated. The solution to this problem is two-fold, the first being a concerted attempt to reduce the visual impact of each visual device, and the second to provide an interactive user interface to provide the viewer with the exact needed information. Such a user interface provides not only the ability to remove unwanted visuals, but also query specific data values, data attributes such as number of observations, and statistical measures.

4 The Summary Plot for 1D Categorical Data

The Summary Plot (Figure 1 is a visual device designed to describe the distribution of a data set including interesting statistical measures as well as tools for distribution fitting and the display of fea-

*e-mail: kpotter@cs.utah.edu

ture statistics on canonical distributions. The Summary Plot incorporates the Box Plot, Histogram, Moment Plot, and distribution fitting into a unified display. Each piece of the Summary Plot is designed to work cooperatively with all other elements, and the addition of a user interface provides the inclusion of as much or as little of the summary elements as desired.

The first piece of the Summary Plot is the *Abbreviated Box Plot*, a refinement of the traditional Box Plot which minimizes the visual impact and maintains the data signature. In this implementation, the minimum and maximum values (including extreme outliers), upper and lower quartiles, and mean are represented. The box surrounding the inner quartile range is reduced to its corners, and the line representing the mean is extended past the edge of the box, to allow the mean to be easily seen, even when the quartiles are spatially close together. While the traditional Box Plot would not be considered visually cluttered, this reduction emphasizes the values expressed by the plot and leaves room for additional visual information.

The histogram is a technique which manageably describes the density of the distribution by binning the data observations. The histogram is added to the Summary Plot as a collection of quadrilaterals, their height reflecting the bin size, and their width indicating density. In addition, density is encoded through each color channel independently, the red channel being normalized log density, the green channel, square root of normalized log density, and the blue channel, normalized linear density. Each of the color channels give a distinct insight to the density and combine to the histogram color seen in Figure 1. Presenting the density of a distribution along with the Box Plot can provide a more clear summary of the data, for example masses of data falling outside the inner quartile range, modality, and the skew, or heaviness, of the data on one side of the mean.

Central moments can be used to reinforce and expand on the statistics of the Box Plot and histogram, and include measures such as mean, standard deviation, skew, and kurtosis or peakiness. The Moment Plot uses glyphs to indicate the location of the mean, and the values of the higher order moments away from the mean. Each glyph is designed to visually express the meaning of that moment, for instance the mean is a small cross that aligns with the median when they are equal (i.e. a normal distribution), and standard deviation is conveyed as brackets falling on either side of the mean. In addition, similar brackets are placed 2 standard deviations away from the mean, giving a sense of the outliers of the data. While the use of some of the higher-order central moments are problematic in that their calculations can be unstable, they do provide some insight to the distribution of the data. Work currently in progress is to determine the most stable methods of calculating the central moments, as well as investigating other descriptive statistics.

Often, understanding the characteristics of a particular data set is less interesting than determining the canonical distribution that best fits the data. This is due to the fact that the feature characteristics of the canonical distributions (such as Normal, Poisson, and Uniform distributions) are well known. The final element of the Summary Plot is a Distribution Fit Plot which is either the best fitting distribution in a library of common distributions, or a distribution chosen by a user. This fit distribution is displayed symmetrically as a dotted line showing the density of the distribution along the axis. Through the user interface mentioned above, the user can also get information about the parameters of the fit distribution, as well as closeness-of-fit statistics.

The resulting Summary Plot provides a simple overview of the data, and techniques for further data exploration and exact statistical measures. The successfulness of this plot will be used as inspiration for higher dimensional data sets.

5 Expanding this work to higher dimensional data

The main challenges to be addressed when expanding the Summary Plot to higher dimensional data sets are finding a visual presentation that is simple and integrates into the data display, and finding visual metaphors such as those in the Moment Plot that are meaningful in 2 and 3 dimensions. The approach that will be taken to accomplish this will be to use transparent surfaces, contours and silhouettes to indicate ranges as well as specific values within the data display. A user interface will be important in allowing a variety of statistics to be used, and to ensure that the display of these measures does detract from existing visualization methods.

A first attempt at such a visualization can be seen in Figure 2. In this figure, uncertainty information of a data simulation of the electric potentials of the heart are shown as transparent surfaces. The surface representing the mean uses the direct mean values, since they have a spatial unit. The position of each of the other surfaces is relative to the mean surface. The values of each of the moments can be seen in the grayscale images at the bottom. While this visualization is not a full proof of concept for the proposed work, it does demonstrate the use of transparent surfaces to express the spatial locations of statistical measures commonly used to describe uncertainty.

The work completed to date includes the development of the Summary Plot and its user interface. Work on the two-dimensional data set of the electric potentials is in progress, and exploration of three-dimensional data sets will begin late fall.

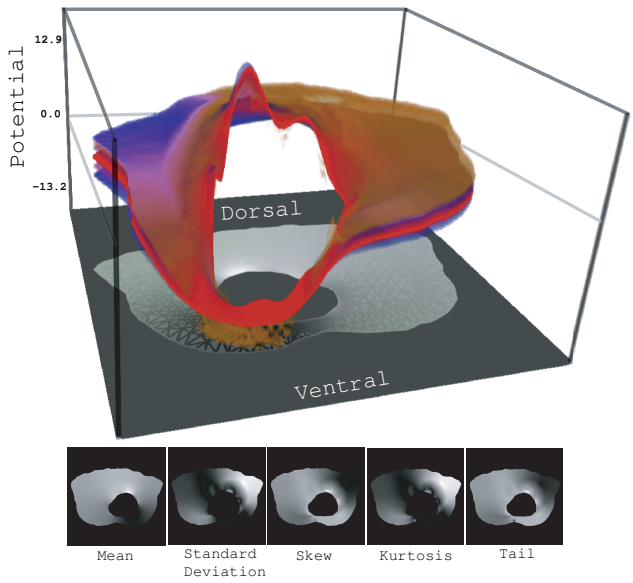

Figure 2: Uncertainty Information for Electrophysiological Potential Models of the Heart. Data provided by S. Geneser and R.M. Kirby [1]

Acknowledgements

*This work was supported in part by ARO(DAAD19-01-1-0013), NSF 03-12479 and ARO grant W911NF-05-1-0395.

References

[1] S. E. Geneser, R. M. Kirby, and F. B. Sachse. Sensitivity analysis of cardiac electrophysiological models using polynomial chaos. In *Proceedings of the 27th Annual IEEE EMBS*, 2005.

[2] C. R. Johnson and A. R. Sanderson. A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications*, 23(5):6–10, 2003.

[3] A. Pang, C. Wittenbrink, and S. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, Nov 1997.