Visual Summary Statistics
UUCS-07-004

Kristin Potter[*] Joe Kniss[†] Richard Riesenfeld[‡]

University of Utah

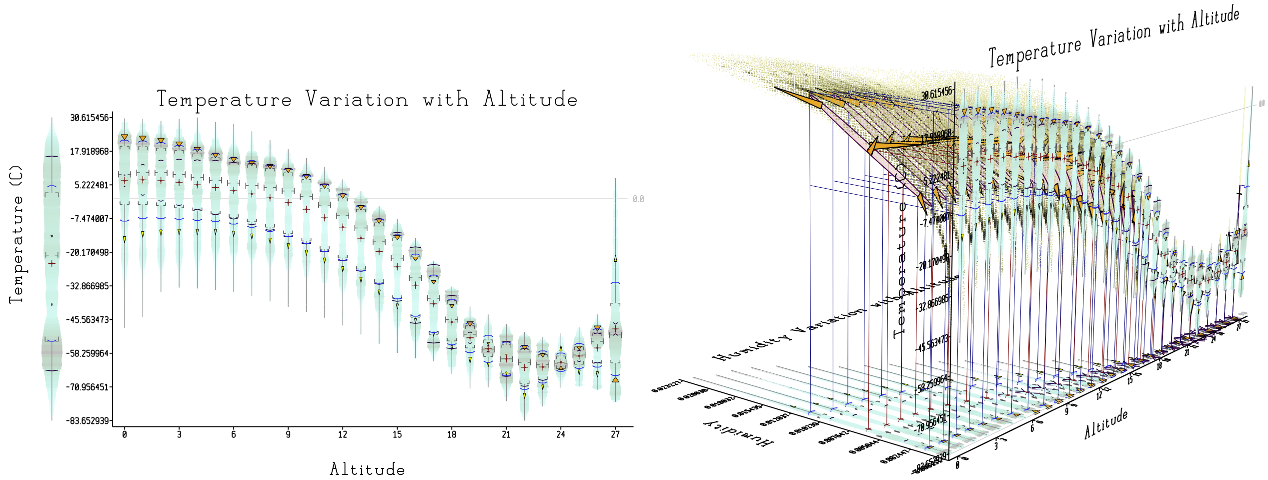


Figure 1: (Left) Summary plot of temperature data and (Right) covariance information between temperature and humidity data.

Abstract

Traditionally, statistical summaries of categorical data often have been visualized using graphical plots of central moments (e.g., mean and standard deviation), or cumulants (e.g., median and quartiles) by box plots. In this work we reexamine the box plot and its relatives and develop a new hybrid summary plot that combines moment, cumulant, and density information. In view of the important role of plots in decision making, our work focuses on incorporating additional descriptive parameters while simultaneously improving the comprehensibility of the summary plots using advanced visual techniques. In many complex situations providing a comprehensive view of the data requires additional summary characteristics, therefore, we submit that these additional parameters, like higher-order central moments can be valuable elements of multi-dimensional summary displays.

CR Categories: G.3 [Probability and Statistics]: Multivariate Statistics— [I.6.9]: Visualization—Information Visualization, Visualization Techniques and Methodologies

Keywords: Summary Statistics and Plots

1 Introduction

As the sophistication of scientific simulation and measurement devices increases, so too does the quantity of data generated. In recent years, we have witnessed an unprecedented demand for the visual analysis of ever higher resolution datasets. Examples include large-scale simulations of important physical phenomena and 3D radiological scans of the human body. A clear understanding of important characteristics in the data through direct inspection is not practical. As such, summary techniques play an important role in data analysis by extracting salient features or descriptors, which can then be presented graphically. This distillation of the data allows the scientist or decision maker to understand and interpret the essential structure data by providing a direct visual, yet quantitative comparison of categories and a global overview of the entire dataset.

Cumulant statistics, such as median and quartile values, are among the most commonly used summary statistics. These statistics partition the data into equally sized groups, revealing insights into the layout of the data such as the range of values and where the majority of the data lie. While these quantities are important, they do not provide information about more subtle yet equally important characteristics of the data set such as whether the data is peaky or skewed. Higher order moments, however, do indicate this kind of information and thus are useful in summary plots. Additionally, density information, if available, should be included with a data summarization.

Creating a summary plot style that clearly conveys essential structures is difficult when additional information is included. Typically, the box plot [15] is used to convey the quartile range of a data set. The principal advantage of the box plot is its elegant simplicity of design. Overlaying large amounts of information on top of the box plot leads to visual clutter that diminishes the effectiveness of the summary. In this work, we have abbreviated the box plot and created *moment glyphs* designed to reduce visual clutter while staying highly informative. The presentation relies on the presence of redundant visual information to reinforce interpretations as well as ensure that our presentation method remains informative, even if some of the statistical modalities are missing. Our goal is to create

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

[†] e-mail: jmk@sci.utah.edu

[‡] e-mail: rfr@cs.utah.edu

a highly informative summary plot that maintains the aesthetic appeal of the box plot while introducing additional parameters which provide insight into the data distribution.

In order to create effective visualizations of data summaries several challenges must be addressed. The first is understanding how statistically to abstract and summarize the data. Typically, cumulant information is used to express a data summary. While this type of summary expresses important information about a distribution, these values alone may not be enough. For example, one may come across two distributions, one uni-modal (having one data value occurring most frequently) and the other multimodal (multiple most frequent values) whose box plot signatures are the same. Investigating just the box plot could lead to the erroneous assumption that the two distributions are very similar. The addition of higher order moment statistics reveals not only distinctions in modality, but other characteristics of the distribution such as skew and peakiness. In this work, we seek to create a foundation for understanding how these statistical modalities can work to create an effective data summary.

The presentation of density and moment information as an augmentation of the box plot can increase the information content of the plot while maintaining its concise form. The box plot has a canonical feel; the "signature" of the plot is easily recognizable and does not need much explanation to allow for a full understanding. Our goal is to create a summary plot that incorporates higher order information smoothly with the box plot in hopes that the summary plot will similarly develop into an easily recognizable signature of the data summary.

2 Background

The main goal of this work is to present summary statistics in a concise, informative manner while conveying the greatest amount of information about the underlying data distribution as possible. As such, previous work from data visualization and statistical techniques for graphical presentation is examined.

2.1 Graphing Principles and Techniques

Creating graphics for data presentation is a difficult task involving not only decisions about data display but also data interpretation. Often, the graphic is intended to show specific characteristics of the data, and the presentation style should make this intended purpose clear. Poor presentation style can be distracting or even mislead the viewer to erroneous conclusions. To alleviate these situations, design practices for effective data visualization are outlined in numerous sources [16, 6, 14]. These references not only direct the scientist towards the "correct" graphical technique for data types, but also describe how a visualization is interpreted by the viewer and suggest methodologies to influence this interpretation.

2.2 Statistical Plotting Techniques

One of the most common approaches to graphing summary statistics is the box plot [15] (or range bar [12]). A variety of box plots can be seen in Figure 4. Typically, the box plot is used to divide the data into four equally sized groups by drawing a box that extends from the upper to the lower quartile, and dividing this box by a line at the median. Lines (or "whiskers") locate the minimum and maximum values in relation to the quartile range and outliers can be indicated with an open circle. This approach is an effective method for quickly summarizing and comparing data distributions.

The box plot can also used to show additional information beyond the five number summary. A survey of the introduction and evolution of the box plot can be found in [5]. The variations of the box plot range from simply changing the width or notching to describe population sizes or confidence [10] to more extreme modifications to express density [9, 4], modality, or multivariate data summaries [11, 8, 13].

A variation of the box plot most closely related to the work presented here enhances the traditional plot by thickening the quartile lines to express skew, modality, and kurtosis [5]. While this approach is straightforward and clean, we desire a representation for these values that has greater visual impact and a more intuitive interpretation.

2.3 From Plots Towards Better Data Comprehension

While a box plot is effective in expressing information about a single data distribution, more complex data sets require methods for navigating the data space to achieve more a complete data understanding. Brushing [3] is a technique that allows the user to select categories of data and see the correlation of the remaining data set. Similarly, the contour spectrum [2] plots an assortment of metrics to provide a quantitative understanding of the data, and allows the user to select specific values for the metric variables which are reflected in the plot and guide the user towards relevant visualizations.

3 The 1D Summary Plot

The main challenges encountered in creating the hybrid summary plot involve creating visual metaphors which encourage a meaningful interpretation of the data. While the meanings of summary information are well known, the visual presentation of this information has yet to be completely described in an effective manner. Extensive previous work has produced a generally universal treatment of cumulant summary information in a clean, concise manner. The box plot has been refined numerous times, resulting in a highly effective style of presentation, including the incorporation of additional information such as density. Our goal is to maintain the clean style introduced by the box plot while increasing the amount of summary information.

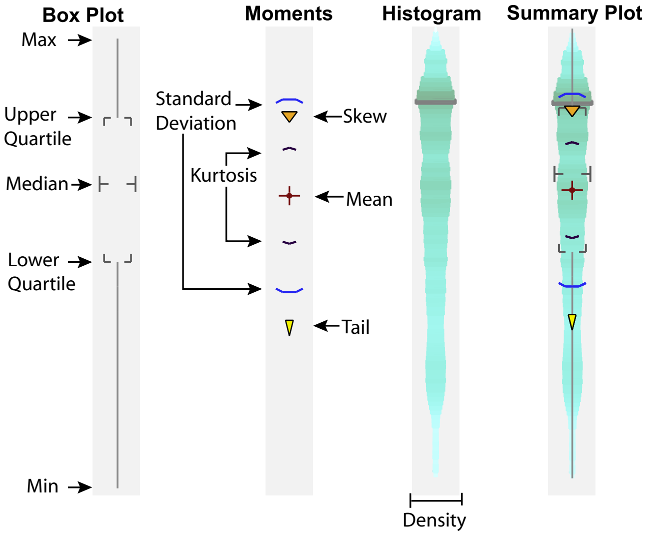


Figure 2: Anatomy of a Summary Plot

The hybrid box plot that we are introducing can more formally be titled the *summary plot*. In this display not only is the quartile information present in the form of a slightly modified box plot, but also a collection of moments and density information. The anatomy of a summary plot can be see in Figure 2. As shown in this figure, we use an abbreviated form of the traditional box plot to convey the minimum and maximum values, upper and lower middle quartiles and the median. Each of the central moments is expressed as a glyph, the design of which reflect the semantic meaning of the moment. Finally, a histogram is added to convey the density of the distribution.

3.1 Quartiles and the Histogram

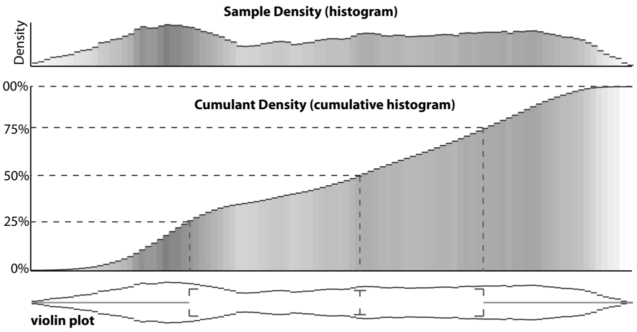


Figure 3: Histogram (top), cumulant histogram (middle), and violin plot (bottom).

One of the simplest ways to describe a data distribution is to calculate the *quartiles* of the data set. A quartile partitions the ordered data into four equally sized subsets such that 25% of the data is less than the lowest quartile, 50% of the data is less than the next quartile (*i.e.* the median) and 75% of the data is less than the highest quartile. There are conflicting conventions concerning whether the term "quartile" refers to the specific data value that cuts off the partition or the subset. In this paper we adopt the former definition in order to be able spatially to place the quartile values. Figure 3 demonstrates this distinction, plotting a single data distribution as density and cumulative histograms. At the top of the figure, a histogram is displayed; the height at each point reflects the density of the distribution at that data value. Below is a cumulant histogram in which density is successively added. Reference lines illustrate the quartile partitioning. At the bottom is a violin plot [9] displaying both the distribution density and cumulant information in a compact form.

The calculation of the cumulant quartile is based on the histogram. The histogram employs a user specified number of bins to sort the data based on value, giving a rough estimate of the density of the data distribution. From the histogram, the quartile values are found using a straightforward counting algorithm. The position of the quartile value is determined by dividing the number of data points by the desired quartile position, and counting the sorted data in the histogram until the quartile position is reached.

The traditional approach to presenting quartile information is through the box plot. Using this technique, a box is drawn around the inter-quartile range (the range between the upper and lower middle quartiles), the median position is denoted by a line through the box, and lines extend to the minimum and maximum values. In efforts to maximize the ratio of information to ink consumption and improve aesthetics, the box plot has been refined numerous times. Four versions of the box plot can be seen in Figure 4. The topmost plot is the range bar, invented by Mary Eleanor Spear [12], next is John Tukey's box plot [15], Edward Tufte's quartile plot [14], and finally our abbreviated box plot. Our plot closely resembles Tukey's box plot with a few distinctions. First, the edges of the box have been removed along with the center of the median and quartile lines. The motivating factor in this change is to reduce the visual clutter that occurs when moment and density information is overlaid with the box plot. Additionally, the median lines are extended slightly outside the boundary of the box, emphasizing the position of the median and insuring this position does not get lost with the addition of more information.

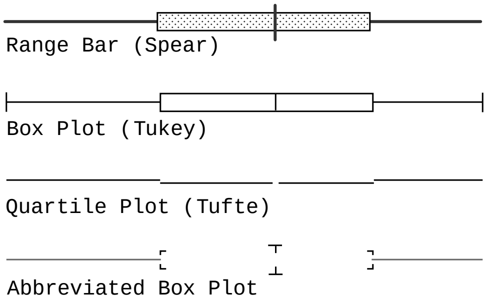


Figure 4: Variations of the box plot. From top to bottom: the range bar [12], the box plot [15], the quartile plot [14], and our abbreviated box plot.

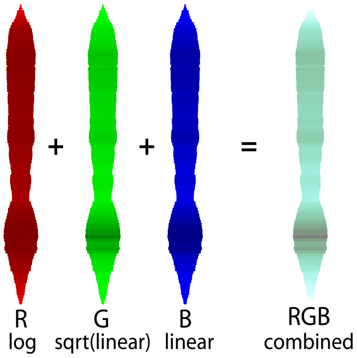


Figure 5: The three color channels of the histogram color map.

In addition to summarizing the distribution of the data through the box plot, the density information itself can be added to the visualization in the form of a histogram. This is similar to the violin plot [9] in that we show the density amount by varying the width of the quadrilaterals used to represent the bins of the histogram. Additionally, the histogram is color mapped based on the density. This color mapping is a redundant mapping combining the three color maps shown in Figure 5. On the left (red) the color map is the normalized log density. Next, (green) the color map is square root of the normalized density and finally, (blue) normalized linear density. While each of these encodings can stand alone, we preferred the redundant encoding due to the fact that the darkest stripes appeared in the areas of the highest density and the resulting color is visually pleasing.

A principal goal of this work is to summarize the distribution of a data set. The histogram is an estimation of that distribution and while its presentation with moment plots is redundant, we can imagine a situation in which we do not have the data distribution but are given only summary data. Thus, the summary display should not only reiterate the distribution when presented with the histogram, but also be able to convey it independently.

3.2 Moments

The moments of a distribution are statistical measures of certain characteristics, the most well known moments being mean and standard deviation. The main distinction between the summaries presented by the quartiles and the moments is that the quartiles give information about the location and variation changes in the data, while moments express specific characteristics of the distribution such as “peakiness”. One of the drawbacks of using only a box plot to summarize a distribution is that multiple, distinct distributions can have the same box plot signature; for instance a bimodal and uni-modal distribution could have identical quartiles. Adding moment information exposes these types of distinctions while maintaining the simplicity of the quartile summary.

The following is a list of the equations used to calculate the various moments, as well as the notation that will be used throughout the paper:

Given a data set $\{x_i\}_{i=0}^{N}$, we define the following quantities:

Expected Value of x:	$< x >$
Central Moments:	$\mu_k \simeq \frac{1}{N}\sum_{i=0}^{N}(x_i - \mu_1)^k$
Mean:	$\mu_1 \simeq \frac{1}{N}\sum_{i=0}^{N} x_i$
Variance:	$\mu_2 \simeq \frac{1}{N}\sum_{i=0}^{N}(x_i - \mu_1)^2$
Standard Deviation:	$\sigma = \sqrt{\mu_2}$
Skew:	$\gamma = \frac{\mu_3}{\sigma^3}$
Kurtosis:	$\kappa = \frac{\mu_4}{\sigma^4}$
Excess Kurtosis:	$\kappa_e = \frac{\mu_4}{\sigma^4} - 3$
Tail:	$\tau \simeq \frac{1}{N}\sum_{i=0}^{N}(x_i - \mu_1)^5$

where N is the number of data samples.

Table 1: Moment Notation and Equations

A valuable way to gain intuition into how moments express characteristics of a data distribution comes from the use of moments in physics (Figure 6). In this example, a beam is placed on a fulcrum, the position of which is dictated by the mean [1]. The moments can then be thought of as weights used to balance the beam, each moment having a specific role in dynamically balancing the system. While this approach is not meant to be a physically based explanation of moments, those unfamiliar with the role of moments in statistics may find this abstraction helpful.

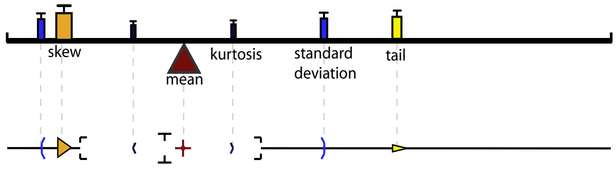


Figure 6: Moment Arm Abstraction

3.2.1 Mean, Variance and Standard Deviation

The most familiar and frequently used moments are mean and variance (the first and second moments). The average of the data values is an estimator of the mean of the underlying distribution, or the expected value of a random variable. Variance is a measure of the dispersion of the data indicating the distance a random variable is likely to fall from the expected value. Standard deviation is simply the square root of variance. For the summary plots, we use only mean and standard deviation, as standard deviation is derived from variance.

The addition of mean and standard deviation into the summary plot is very straightforward. The mean is rendered as a dark red cross with a small circle in the center, denoted by a stylized bull’s eye. The width of the lines making up the cross are constructed so that when the mean and median are displayed at the same location, the glyphs line up, forming a straight line across the plot. This emphasises normal distributions, and quickly reveals when a distribution varies from a Gaussian. A close up of this can be seen in Figure 7. Standard deviation is rendered as two glyphs on the plot, as are all even moments. Two blue curved lines are placed on either side of the mean to express the average variation from the mean.

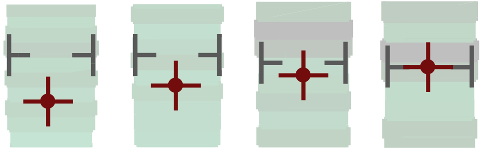


Figure 7: Mean and median glyphs align when values are equal.

3.2.2 Skew

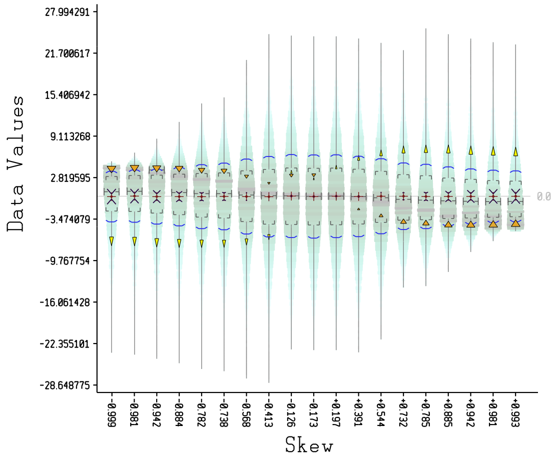


Figure 8: Distributions with small (left) and large skew (right).

Skew is a measure of the degree of asymmetry of a distribution; that is, the amount that the data is pushed to one side or the other. Figure 8 shows various distributions with skew varying from small to large. Based on the balance beam abstraction (see section 3.2), we use a large triangle to denote skew in the summary plot and place it so that it rests on the end of the distribution with the most weight and pointing at the tail. Mathematically, we calculate the placement of the skew glyph by first finding skew as defined in Table 1 and placing the glyph $-\gamma$ distance away from the mean, with the apex of the triangle pointing toward the tail of the distribution.

3.2.3 Kurtosis

Kurtosis is a measure of how peaked or flat topped a distribution is compared to a normal (Gaussian) distribution. Excess kurtosis is the standard kurtosis measure normalized by the kurtosis of a Gaussian. An example of distributions with different kurtosis can be seen in Figure 9 where a flat, box-like distribution can be seen on the far left. This type of distribution has large, negative kurtosis (*i.e.* $\kappa_e < 0$) and is called *platykurtic*. Moving right, the kurtosis values increase, getting very close to a *mesokurtic* (normal) distribution (*i.e.* $\kappa_e = 0$) and moving on to a highly peaked, *leptokurtic* (*i.e.* $\kappa_e > 0$) distribution.

The glyphs chosen to represent kurtosis reflect the aforementioned categories of kurtosis. The glyphs are rendered using a deep

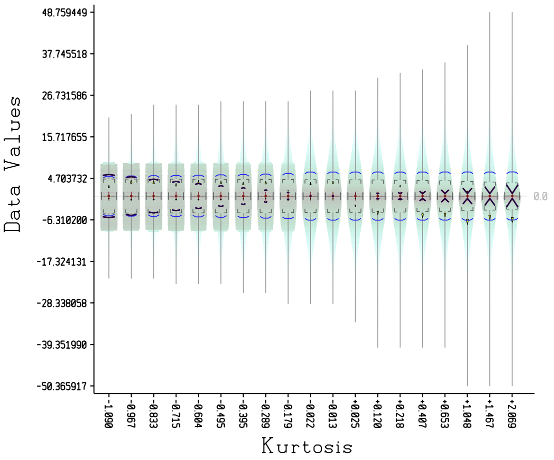


Figure 9: Distributions with small (left) to large (right) kurtosis.

purple color and are scaled so that their size reflects their magnitude away from 0. To distinguish between flat and peaked, the glyphs assume a flat or sharp shape depending on the sign of kurtosis. Thus, for a highly positive value, the glyph is very pointy, and the more negative the kurtosis value, the flatter the glyph.

3.2.4 Tail

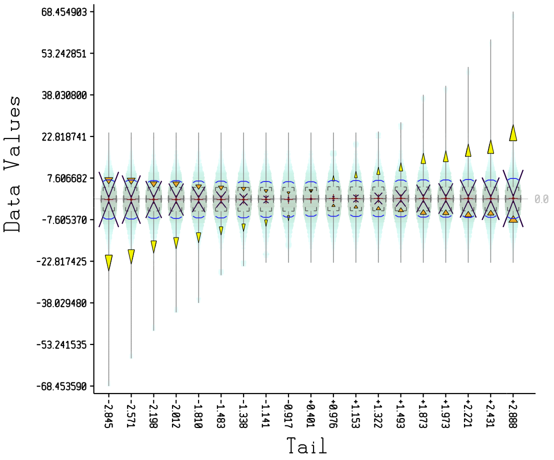


Figure 10: Distributions with small (left) and large (right) tail values.

The final moment that we add to the summary plot is what we refer to as tail, which is based on the fifth central moment μ_5. The quantity is sensitive to distribution asymmetry farther way from the mean when compared to the skew. Tail will have a high magnitude when there are additional modes in the distribution or strong outliers. Like skew, tail is rendered as a triangle pointing in the direction of asymmetry. However, unlike skew, tail is rendered on the same side of the mean as its sign. The tail glyph is rendered as a sharp arrow head, where both the size and sharpness is dependent on the tail quantity. The visual effect of this glyph should indicate that there are a significant number of samples biased far from the mean. Figure 10 shows a set of distributions which have tail values varying from very negative to very positive. The center distribution is a Gaussian.

4 Joint 2D Summaries

While a statistical summary for a 1D categorical data set is highly useful, methods for comparing multiple, correlated data sets are necessary to understand how samples with multiple distinct data values are related. In this section we explore methods for summarizing categorical data with pairs of values associated with each sample. Note that we drop the summary of cumulants for higher dimensional distributions. We do feel, however, that the cumulant summary is important even in higher dimensions, and there does exist a generalization of the box plot, known as the bag-plot [11]. Unfortunately, the bag plot approach does not necessarily have the same correspondence to cumulant distributions as does the box plot. It is a suitable approximation for many applications, but we will defer discussion of multivariate cumulant summaries to future work.

4.1 Joint Density

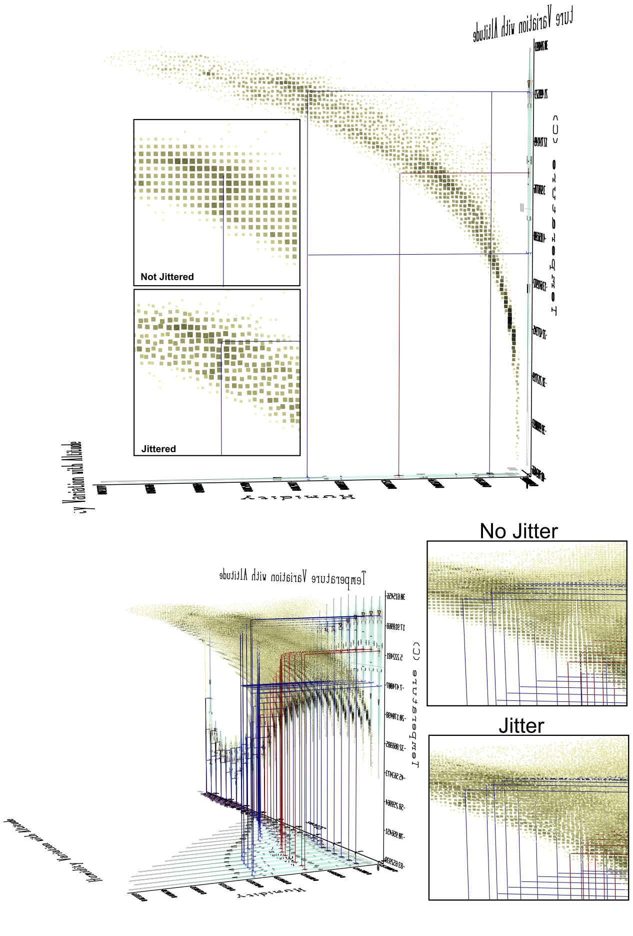


Figure 11: The joint histogram.

The density of a set of samples drawn from a 2D distribution can be directly visualized using a joint histogram. A joint histogram can be generated by subdividing the 2D domain into NxN bins, and, for each sample, incrementing the bin-count indexed by its pair of data values. Our system displays the joint histogram by rendering a quadrilateral at each bin location scaled by the square-root of the normalized density for that bin.

When multiple categories are summarized simultaneously using joint histograms, they tend to produce aliasing artifacts due to the regularity of the bin spacing. To alleviate this problem, we jitter the position of the quadrilaterals for each bin, where the magnitude of the jitter is inversely proportional to the quadrilateral's scale. This constraint ensures that the quadrilateral is drawn at a randomized location, but is always inside the bin.

4.2 Covariance

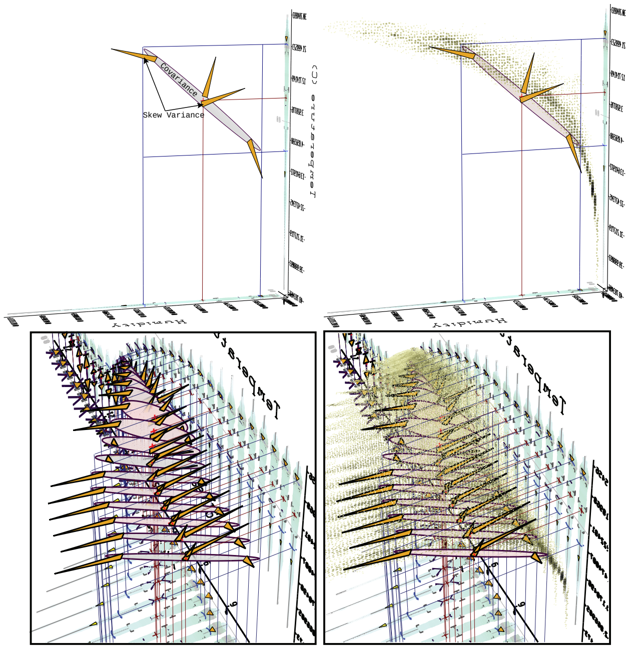


Figure 12: Covariance and Skew Variance

For multivariate distributions, the covariance matrix is the analogue of variance in 1D distributions. The covariance of two data sets, $\{x_i\}_{i=0}^N, \{x_j\}_{j=0}^N$ can be defined by:

$$V_{ij} = < (x_i - \mu_i)(x_j - \mu_j) > = \frac{1}{N-1} \sum_{k=0}^{N} (x_{i_k} - \mu_i)(x_{j_k} - \mu_j)$$

where μ_i and μ_j are the means for each data set. Covariance is a measure of how the two data sets vary in relation to each other. For our presentations, the covariance matrix is used to transform a unit disk, the way in which the disk is stretched visually relates to covariance of the data sets. Since we are actually interested in a multivariate analogue of standard deviation, we scale the covariance ellipse-disk glyph by,

$$\text{scale} = \frac{\sqrt{\text{ev}_{\text{max}}}}{\text{ev}_{\text{max}}},$$

where ev_{max} is the maximum eigenvalue of the covariance matrix. Figure 12 shows the covariance ellipse between the first categories of the temperature and humidity data sets. The covariance glyph is laid on top of the joint histogram, and for reference, the mean and standard deviation of both data sets are extended into the joint space using lines. In this case the 1D summary plots for each of the variables represents the marginalized distribution.

4.2.1 Skew-Variance

Just as covariance is the analogue of variance, higher order multivariate moments can also be described as matrices. The so called "skew-variance" of two data sets, $\{x_i\}_{i=0}^N, \{x_j\}_{j=0}^N$ can be expressed by two matrices, $V_{i^2 j^1}$ and $V_{i^1 j^2}$ where:

$$V_{i^m j^n} = < (x_i - \mu_i)^m (x_j - \mu_j)^n > = \frac{1}{N-1} \sum_{k=0}^{N} (x_{i_k} - \mu_i)^m (x_{j_k} - \mu_j)^n$$

In general, these matrices are neither symmetric, nor positive definite. $V_{i^2 j^1}$ and $V_{i^1 j^2}$ are, however, the transpose of one another; therefore we only need one to capture the information expressed by both. Skew variance is visualized using four sharp arrows pointing in the direction of the skew. These directions are defined by the column vectors of $V_{i^2 j^1}$ and $V_{i^1 j^2}$. As with covariance, skew-variance visualizations are scaled by

$$\text{scale} = \frac{\sqrt[3]{\text{ev}_{\text{max}}}}{\text{ev}_{\text{max}}},$$

where ev_{max} is the maximum eigenvalue of the skew-variance matrix.

4.3 Continuous 2D Categorical Data

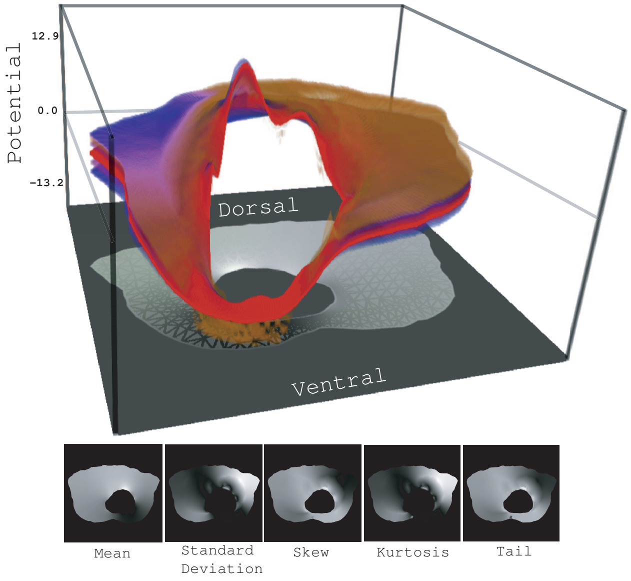


Figure 13: Electric potentials of the heart data.

The previous sections discussed the visual summary of a limited number of categories with 1D and 2D distributions. We now consider the case of a continuous 2D domain with a 1D distribution. One can think of this case as having a continuous 2D array of categories. Our system treats the summary visualization as a height field. Figure 13 shows an example of a stochastic numerical simulation of the electric potential in the human torso [7]. The simulation domain is a 2D second-order finite element mesh. The goal of the simulation is a study of variation in electric potential induced by perturbations of conductivity in various organs. To gather the data, the simulation was run 100,000 times with different random perturbations of lung conductivity. That is, we have 100,000 samples from the stochastic solutions distribution for each element in the domain. Our summary of the data presents the mean (red), standard deviation (blue), skew (yellow-orange), and kurtosis (purple). Just as with the 1D summaries, we use $\mu_1 - \gamma$ to place the skew, indicating the “heavy” side of the distribution. At every position in this data set excess kurtosis is less than or equal to 0, indicating a flat distribution. The opacity for all higher-order moments is proportional to their magnitudes, *i.e.* they are only visible if they are significant. This scale term is identical to the one used to scale the size of these glyphs in the 1D summary plots. A flat image of the domain colored by mean potential is mapped below the height field summary for reference.

5 DISCUSSION

While the box plot has been used, almost universally, to summarize statistical data for nearly 60 years, there are many characteristics of a distribution that it cannot express. The mean or expected value of the distribution, for instance, is one such characteristic. Without this moment in the summary, a user may incorrectly assume that the median and mean are the same or closely correlated. Certainly, the same can be said about summaries based solely on the moments. This is especially true when the only moments considered are the first two, mean and variance. Such a summary would imply a symmetric uni-modal distribution, like a Gaussian. We have not frequently seen normal distributions summarizing arbitrary distribution data in scientific simulation and imaging. Together, box plot, cumulant and moment summaries express different, yet complementary aspects of the data. However, they may still fail to expose important subtleties of the distribution. The density plots or histograms simply summarize the the data itself. While the histogram summary makes the modes of the data easily discernible, it does not allow the user to predict the median or mean values. By combining all three summary methods, we can feel more confident in the analysis of the data and the questions that the summary is intended to help answer.

The display system for our summary plots was implemented using OpenGL, which allows a user to interactively explore multidimensional summaries. Interactive control of view point and summary content is essential when the goal is to compare many categories with multivariate distributions, *i.e.* joint summaries, or mutidimensional categories, *e.g.* continuous 2D categorical data. The depth complexity of such summaries can be overwhelming for arbitrary viewpoints in static 2D images. When the goal is a 2D image, interactive control allows the user to select view points that focus on key aspects of the summary yet include essential context. The generation of summary snapshots utilises “gl2ps”, a freely available library for directly converting OpenGL images to Encapsulated PostScript (EPS). This mechanism allows us to preserve the resolution independent characteristics of the original vector art.

The assembly of our summary plots also emphasizes the concept of marginal summary. As seen in Figure 1 (left), we show summaries for each category, but add an additional summary that covers all categories to the left of the value scale. This “marginal” summary expresses global data characteristics that may not be obvious in the individual category plots. This concept extends easily to joint distributions, where the 1D summaries for each value become a marginal summary for the 2D distribution, as seen in Figure 1 (right).

The development of this work is driven by the needs of large-scale simulation and medical image analysis. This data generally has on the order of millions of samples per category. The kind of summary generated for this data is extremely robust; millions of samples are generally sufficient for reliable moment and density estimation. When the number of samples available for each category is substantially fewer, the measurement of higher-order moments can break down. These moments, *e.g.* skew, kurtosis, and tail, can be extremely sensitive to outliers when there are not enough samples to adequately characterize the underlying distribution. The density plot (histogram) visualization becomes extremely important in this case. This aspect of the visualization should make it readily apparent to the user that the summary is based on a sparse number of samples. Our concern here motivates further research on incorporating measures of sufficient statistics as part of the summary.

Our future work will focus on further generalizations of summary plots to higher dimensions, both in terms of multivariate distributions and multi-dimensional category domains. We are also interested in the automatic detection of distribution characteristics such as multi-modality and correlating with specific analytic distributions (*e.g.* Normal, Poisson, Raleigh, Chi-squared, etc...). Our log-term goal is the development and release of a fully interactive system for summary visualization that provides direct access to the summary process, which will utimately allow interactive summaries embedded in electronic report documents.

6 CONCLUSION

The box plot is a highly effective means for conveying cumulant summary statistics. Using the box plot as inspiration, we have created a hybrid summary plot that incorporates cumulant statistics, density, and high-order moments. We have demonstrated a generalized approach for provide joint 1D comparisons as well as summaries of 2D categorical data. Our system aims at reducing visual clutter, while redundantly encoding information and simultaneously presenting a large amount of data as a visual signature. The presentation of data in a summarized and easy to read form can quickly communicate large amounts of data, emphasize meaningful characteristics, and facilitate visual comparisons.

REFERENCES

[1] Lee J. Bain and Max Engelhardt. *Introduction to Probability and Mathematical Statistics*. Duxbury Press, 1992.

[2] Chandrajit L. Bajaj, Valerio Pascucci, and Daniel R. Schikore. The contour spectrum. In *IEEE VIS ’97: Proceedings of the 8th conference on Visualization ’97*, pages 167–173, 1997.

[3] Richard A. Becker and William S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, May 1987.

[4] Yoav Benjamini. Opening the box of a boxplot. *The American Statistician*, 42(4):257–262, November 1988.

[5] Chamnein Choonpradub and Don McNeil. Can the box plot be improved? *Songklanakarin Journal of Science and Technology*, 27(3):649–657, 2005.

[6] William S. Cleveland. *The Elements of Graphing Data*. Hobart Press, 1994.

[7] Sarah E. Geneser, Robert M. Kirby, and Frank B. Sachse. Sensitivity analysis of cardiac electrophysiological models using polynomial chaos. In *Proceedings of the 27th Annual IEEE EMBS*, 2005.

[8] Kenneth M. Goldberg and Boris Iglewicz. Bivariate extensions of the boxplot. *Technometrics*, 34(3):307–320, August 1992.

[9] Jerry L. Hintze and Ray D. Nelson. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184, May 1998.

[10] Robert McGill, John W. Tukey, and Wayne A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, February 1978.

[11] Peter J. Rousseeuw, Ida Ruts, and John W. Tukey. The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):382–287, November 1999.

[12] Mary Eleanor Spear. *Charting Statistics*. McGraw-Hill, 1952.

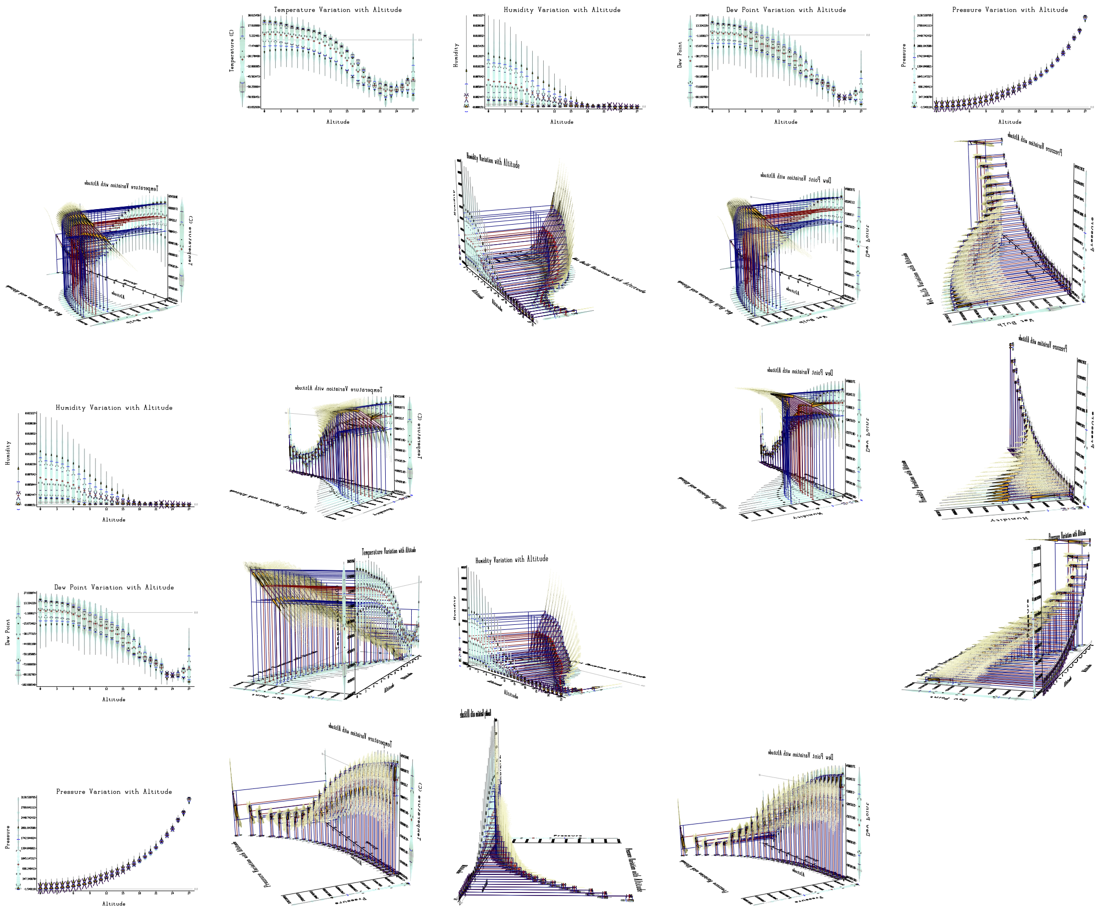
[13] Phattrawan Tongkumchum. Two-dimensional box plot. *Songklanakarin Journal of Science and Technology*, 27(4):859–866, 2005.

[14] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[15] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[16] Leland Wilkinson. *The Grammar of Graphics*. Springer-Verlag New York, Inc., 1999.

Matrix of Weather Data


Temperature Variation with Altitude
Temperature (C)
Altitude
Humidity Variation with Altitude
Humidity
Altitude
Dew Point Variation with Altitude
Dew Point
Altitude
Pressure Variation with Altitude
Pressure
Altitude
Humidity Variation with Altitude
Humidity
Altitude
Dew Point Variation with Altitude
Dew Point
Altitude
Pressure Variation with Altitude
Pressure
Altitude