Shahid Sadoughi University of Medical Sciences School of Public Health

Journal of Social Behavior and Community Health

Journal of Social Behavior & Community Health

Social Determinants of Health Research Center

Violin Plots: An Enhanced Tool for Data Visualization in Health Studies

Zohreh Shishebor [a], **Zahra Sajjadnia** [a*], **Maryam Sharafi** [a]

Department of Statistics, College of Science, Shiraz University, Shiraz, Iran

ARTICLEINFO

Commentary

Article History:
Received: 11 December 2024
Revised: 26 February 2025
Accepted: 23 March 2025

***Corresponding Author:**
Zahra Sajjadnia

Email:
sajjadnia@shirazu.ac.ir

Tel: +098 09173085288

Keywords: Box Plot, Error Box Plot, Violin Plot, Whisker, Kernel Density Estimation

Citation:
Z Shishebor, Z Sajjadnia, M Sharafi. Violin Plots: An Enhanced Tool for Data Visualization in Health Studies. Journal of Social Behavior and Community Health (JSBCH). 2025; 9(1): 1600-1606.

Introduction

Statistical knowledge enables us to utilize appropriate methods for data collection, conduct accurate analyses, and effectively communicate results. Statistics plays a pivotal role in scientific discoveries, data-driven decision-making, and predictive modeling. Statistical plots serve as powerful tools for enhancing data visualization and facilitating data communication. By examining a statistical graph, underlying patterns and relationships within the data become more discernible.

Visualizations are a far more effective way of communicating statistical findings to a non-technical audience. Charts and graphs make complex data relationships much easier to understand than tables of numbers or statistical reports. Even for technical audiences, a well-designed statistical graphics can quickly convey the key insights. When we are dealing with a new dataset and do not have a specific hypothesis to test, visual analysis excels at revealing patterns, outliers, and potential relationships that might not be apparent from numerical summaries alone. Statistical tests assume a certain structure in the data; visual inspection helps determine if that assumption is justified. Statistical graphs like box plots, scatter plots, and histograms can readily highlight outliers that might be missed by statistical tests focused on central tendencies. These outliers might be errors or genuinely interesting data points requiring further investigation. It is crucial to remember that ocular and statistical analysis are complementary, not competing, approaches, so both should be used. Visualizations guide the choice and interpretation of statistical tests, while statistical tests provide a more rigorous quantification of observed patterns in the data. Using both methods together leads to more robust and reliable conclusions (Hullman et al., 2015; Tan & Shi, 2019; Wainer & Thissen, 1981).

[DOI: 10.18502/jsbch.v9i1.18706]

This study investigated three valuable statistical plots that ocularly represent data variability and compare different groups. These graphs include the box plot, error box plot, and violin where the last graph is newer than the others. Violin plots are novel plot similar to box plots, with rotated kernel density plots to each side which integrates density functions with box plots, providing a richer perspective on data distributions than box plots alone. These plots provide a deeper understanding of data distributions than box plots by revealing the full distribution shape, including modality and skewness. This makes them a valuable tool in statistical analysis across diverse fields, particularly in medical research. Their application is widespread in this field; for example, Wang et al. (2023)(Wang et al., 2023) used violin plots to analyze Parkinson's disease data, while Freeman et al. (2023)(Freeman et al., 2023) used this graph to visualize daily patient phenotypic data during hospital admission, and Zhu et al. (2023)(Zhu et al., 2023) used it in gene expression analysis alongside box plots and prognostic curves. Moreover, Karpefors et al. (2024)(Karpefors et al., 2023) introduced the novel Maraca plot, combining violin plots (with nested box plots) and Kaplan-Meier curves for comprehensive visualization of continuous and time-to-event outcomes.

Numerous other publications demonstrate the versatility of violin designs in clinical research, which is beyond the scope of this paper(Oku, 2024). Therefore, we limit ourselves to the mentioned articles.

To stablish the usage of box plot, error box plot, and violin plot, we demonstrate the utility of these plots with a fictitious example, showing how researchers can extract meaningful insights from data. Imagine a clinical trial testing the effectiveness of a new drug (Drug A) compared to a standard treatment (Drug B) for high blood pressure. The primary outcome is the reduction in systolic blood pressure after six weeks.

The box plot was used to visually compare the distribution of systolic blood pressure changes (before vs. after treatment) between the Drug A group and the Drug B group. This plot shows the median, quartiles, and outliers, allowing researchers to see if one group had a greater overall reduction in blood pressure or a different spread of results. For example, if the box plots show a greater reduction in blood pressure for the Drug A group, with a lower median and smaller interquartile range, this suggests the potential effectiveness of this drug. In addition, the error box plot displays the mean blood pressure change for each group along with error bars representing the standard error or confidence intervals. This plot helps to determine if the difference in mean blood pressure reductions between the groups is statistically significant. Finally, the violin plot is useful to visualize the full distribution of blood pressure changes within each group, providing insights into the shape and spread of the data. It would show if the distributions are skewed or have multiple peaks (modality), which could indicate differences in how the treatments affect different individuals. A narrow, symmetrical violin plot with a single peak centered around a larger reduction in blood pressure. This suggests that Drug A consistently leads to significant blood pressure reductions in most individuals where a wider, more skewed violin plot with a lower peak and possibly even a second smaller peak. This might suggest that Drug B is less effective for some individuals and leads to a wider range of outcomes. The empirical density function within a violin plot offers a richer visualization of the data distribution than a simple box plot. This visual information can be crucial for understanding how treatments affect different individuals and identifying potential subgroups within treatment groups. This is particularly relevant for clinical trials, where understanding the nuances of treatment effects is essential for informed decision-making.

The paper is structured as follows: Section 2 offers an overview of the basic principles and key definitions of box plots, error box plots, and violin plots. Section 3 examines the practical use of these plots especially the violin plot in analyzing a real-world dataset, and finally, in the Appendix, the R

[DOI: 10.18502/jsbch.v9i1.18706]

codes used in this paper are presented.

2- Basic Definitions and Principles

In this section, we explain the definition of the desired plots and introduce different parts of them.

2-1- Box plot (Boddy (2009))(Boddy & Smith, 2009)

Definition: A box plot (also known as a box-and-whisker plot) is a standardized way to display the distribution of a dataset. It visually summarizes five key statistics of the data: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. This makes it useful for comparing data distributions across different groups or analyzing the spread and skewness of a single dataset.

2-1-1- Parts of a box plot

- **Box:** The central rectangle of the box plot represents the interquartile range (IQR).
- **Q1 (First Quartile):** The 25th percentile of the data.
- **Q2 (Median):** The 50th percentile of the data (the middle value).
- **Q3 (Third Quartile):** The 75th percentile of the data.
- **Whiskers:** Lines extending from the box. These indicate the range of the data excluding outliers.
- **Lower Whisker:** Typically extends to the smallest data point within 1.5 * IQR of Q1.
- **Upper Whisker:** Typically extends to the largest data point within 1.5 * IQR of Q3.
- **Outliers:** Individual data points that fall outside of the whiskers. They are often marked with dots or asterisks. Outliers can be influential data points that warrant further investigation.

2-1-2- Advantages and disadvantages of box plot

The box plot is straightforward to understand and interpret; even a person without good statistical knowledge can grasp the basic information. Also, this plot provides a quick summary of the data such as showing the center, spread, and potential outliers of the data. The box plot is very useful to compare groups. It allows the researcher to ocularly compare the distribution of data across different categories (e.g., treatment vs. control groups).

Despite the advantages of this plot, it has some disadvantages such as the loss of detailed information. This plot only shows a summary of the data, not individual values. Another disadvantage of it is that it is not ideal for complex distribution and may not be suitable for datasets with multiple peaks, more than one mode, or very skewed distributions.

2-2- Error box plot (Spear (1969))(Spear, 1969)

Another plot that we discuss is the Error box plot.

Definition: An error box plot, also known as a box plot with error bars, is a type of visualization that combines elements of a standard box plot with error bars, which represent the uncertainty or variability of the data. It provides a more comprehensive picture of data distribution and central tendency, including information about the precision of the estimates.

2-2-1- Parts of the error box plot

- **Box:** The central rectangle of the box plot, representing the interquartile range (IQR), remains the same as in a standard box plot.
- **Q1, Median, Q3:** The first quartile (Q1), median (Q2), and third quartile (Q3) are depicted within the box as in a standard box plot.
- **Whiskers:** The whiskers, extending from the box, typically represent the range of the data excluding outliers, similar to a standard box plot.
- **Error bars:** These are the key additions to the standard box plot. Error bars visually represent the uncertainty around a point estimate (usually the mean). They can be depicted in various ways, but some common methods include:

a.Standard error (SE) bars: Extend above and below the mean by the standard error of the mean. These bars represent the precision of the mean estimate. Smaller bars indicate higher precision.

b. Confidence interval (CI) bars: Extend

[DOI: 10.18502/jsbch.v9i1.18706]

above and below the mean by the margin of error of a confidence interval. They indicate the range within which the true population mean is likely to fall with a certain level of confidence (e.g., 95%).

2-2-2- Advantages and disadvantages of the error box plot.

Error box plots help us understand not only the central tendency mean of the data but also the variability around that estimate. It is particularly useful for comparing different groups or treatments. By comparing the means and error bars, we can assess whether the differences between groups are statistically significant or likely due to chance. This plot is widely used in various fields especially in clinical trials to compare treatment effects between different groups. But despite of box plot, understanding the error box plot needs some statistical knowledge and can visually cluttered with multiple groups.

2-3- Violin plot (Molina et. al. (2022))(Molina et al., 2022)

A new plot that with its ability to depict data density and distribution, offers a more comprehensive and informative representation has been introduced by Hintze and Nelson (1998)(Hintze & Nelson, 1998). In this section, we present some useful information about this plot.

Definition: A violin plot is a statistical visualization that combines the features of a box plot and a kernel density estimation plot. It is used to visually represent the distribution of numerical data, particularly when comparing multiple groups or categories. It works as follows:

Kernel density estimation: The violin plot uses a kernel density estimator to estimate the probability density of the data. This creates a smooth curve representing the distribution of the data points, showing the data center and spread.

Mirrored shape: The estimated density is then mirrored on both sides of a central line, creating a symmetrical shape that resembles a violin. The width of the violin plot at any point represents the density of data points at that value.

Box plot features: The central line within the violin plot typically represents the median of the data. The box plot quartiles (25th and 75th percentiles) are sometimes included within the violin shape, and outliers are often shown as individual points.

2-3-1- Advantages and disadvantages of the violin plot:

- **Advantages:**

Violin plots provide a richer representation of the data distribution than box plots, showing the density of data points across the entire range of the variable. They clearly illustrate outliers, which can be helpful in medical research where outlier values might be clinically significant. Violin plots excel at comparing the distributions of multiple groups, highlighting differences in central tendency, spread, and shape. They help researchers identify patterns, trends, and potential anomalies in the data distribution. In essence, violin plots offer a more informative and nuanced visual representation of data distributions compared to traditional box plots or histograms, making them a valuable tool for exploring and interpreting medical data.

- **Disadvantages**

Violin plots have some drawbacks, such as being misleading or inaccurate if the data is not smooth or continuous or if the sample size is too small. Additionally, they can be difficult to read or understand if the audience is unfamiliar with them or the plot is too complex (Tanious & Manolov, 2022).

3- Real Data Example

In this section, we use a real data set from a cross-sectional study to compare and investigate the capabilities of the plots introduced in the previous sections. This cross-sectional study (January–July 2011) involved 208 patients diagnosed with hemorrhoids at Faghihi Hospital, Shiraz University of Medical Sciences, Iran. Patients were randomly assigned to outpatient and inpatient groups. Patient satisfaction was measured using a 21-item questionnaire encompassing eight domains: surgical outcome, treatment, surgical environment, nursing services, pre-operative

[DOI: 10.18502/jsbch.v9i1.18706]

explanations and instructions, appointment scheduling, surgical costs, and waiting time. A 4-point Likert scale (4=perfect, 3=well, 2=moderate, 1=dissatisfied) was used to assess satisfaction within each domain. Individual domain scores were averaged to create a composite satisfaction score for each patient.

Three plots were drawn using R software and the ggplot2 package. The used codes are presented in the appendix.

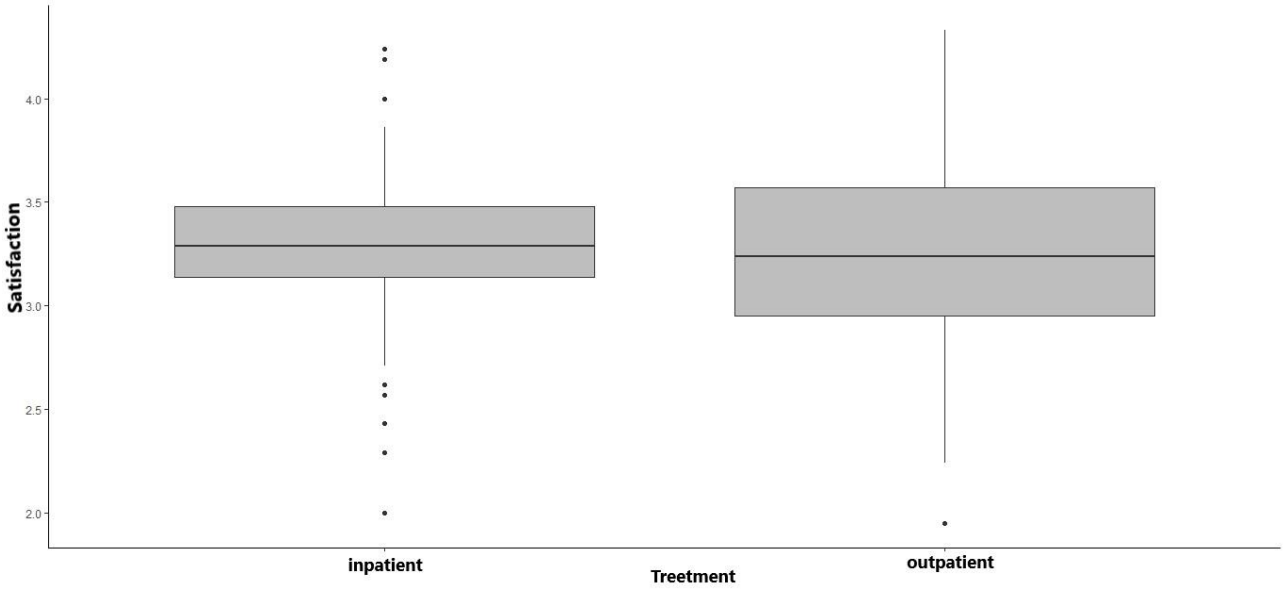

Figure 1. The box plot of the satisfaction of two groups of outpatients and inpatients.

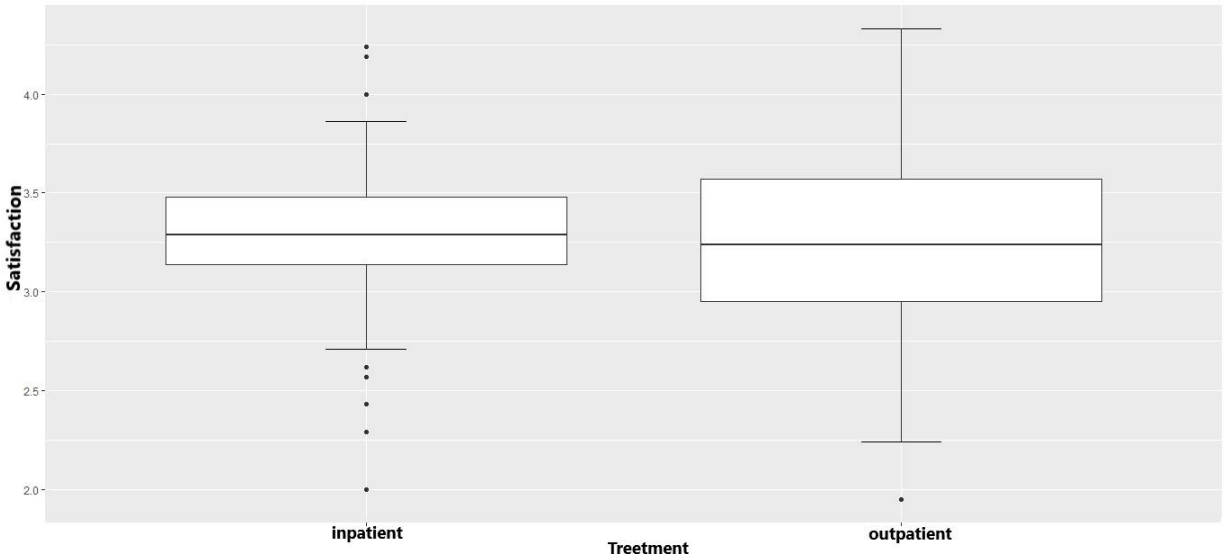

Figure 2. The error box plot of the satisfaction of two groups of outpatients and inpatients.

[DOI: 10.18502/jsbch.v9i1.18706]

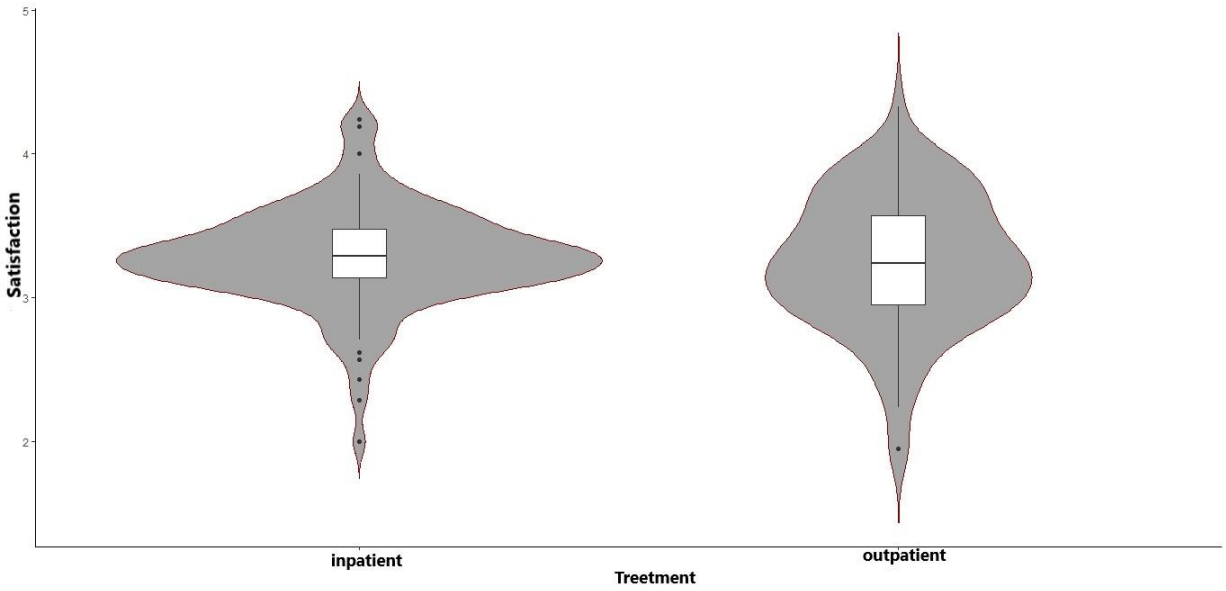

Figure 3. The violin with box plot of the satisfaction of two groups of outpatients and inpatients.

According to the box plot (Figure 1), it can be seen that in dispersion, in the satisfaction scores of inpatients, there are some outliers, but the satisfaction scores of outpatients do not have outliers when these outliers are more specific in Figure 2. The satisfaction scores of inpatients are less dispersed than those of outpatients and the distribution of both data groups appears symmetric. This graph, at first glance, which is usually done superficially, does not provide any information about the mode and its number; however, in Figure 3, especially in outpatients, it can be seen that the data is not symmetrically distributed and it is bimodal. Moreover, the satisfaction scores of inpatients are more concentrated around the median than the satisfaction scores of outpatients. The violin plot combined with the box plot gives us a much more detailed view of the data than the box plot alone. Finally, the violin plot informs us of all the characteristics of the data, while the box plot only deals with the quantiles and their relationship to each other. Therefore, we will lose a lot of information in box plot.

Conclusion

In this paper, we investigated three types of useful plot including box plot, error box plot, and violin plot which are using in the statistical analysis of data specially in the clinical surveys.

As can be seen in the real example, the violin chart along with the box plot is the most efficient of these three plots in analyzing data and expressing all of their distributional features.

References

Boddy, R., & Smith, G. (2009). *Statistical methods in practice: for scientists and technologists*. John Wiley & Sons.

Freeman, R., Noronha, A., & Woods, J. (2023). Next generation phenotyping with quantitative narration for DEGCAGS syndrome. *American Journal of Medical Genetics Part A*, *191*(4), 1020-1025.

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, *52*(2), 181-184.

Hullman, J., Resnick, P., & Adar, E. (2015). Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one*, *10*(11), e0142444.

Karpefors, M., Lindholm, D., & Gasparyan, S. B. (2023). The maraca plot: a novel visualization of hierarchical composite endpoints. *Clinical Trials*, *20*(1), 84-88.

Molina, E., Viale, L., & Vázquez, P. (2022). How

[DOI: 10.18502/jsbch.v9i1.18706]

should we design violin plots? 2022 IEEE 4th workshop on visualization guidelines in research, design, and education (VisGuides),

Oku, M. (2024). Clarinet Plots: Alternative to Violin Plots to Display Zero-inflated Distribution of scRNA-seq Data. *IPSJ Transactions on Bioinformatics*, *17*, 48-54.

Spear, M. E. (1969). *Practical charting techniques*. McGraw-Hill Companies.

Tan, Y., & Shi, Y. (2019). *Data Mining and Big Data: 4th International Conference, DMBD 2019, Chiang Mai, Thailand, July 26–30, 2019, Proceedings* (Vol. 1071). Springer.

Tanious, R., & Manolov, R. (2022). Violin plots as visual tools in the meta-analysis of single-case experimental designs. *Methodology*, *18*(3), 221-238.

Wainer, H., & Thissen, D. (1981). Graphical data analysis. *Annual review of psychology*, *32*(1), 191-241.

Wang, H., Dou, S., Wang, C., Gao, W., Cheng, B., & Yan, F. (2023). Identification and Experimental Validation of Parkinson's Disease with Major Depressive Disorder Common Genes. *Molecular Neurobiology*, *60*(10), 6092-6108.

Zhu, H., Lin, Q., Gao, X., & Huang, X. (2023). Identification of the hub genes associated with prostate cancer tumorigenesis. *Frontiers in Oncology*, *13*, 1168772.

Appendix: R codes of real data

```
#------ Data
 data=read.csv("D:\\violin1.csv",header=T)
 data$treetment <- as.factor(data$treetment)
 install.packages("ggplot2")
 library(ggplot2)

#----- Box Plot
  ggplot(data, aes(x=treetment, y=satisfaction)) +
   geom_boxplot(fill="gray")+
   labs(title="",x="Treatment", y = "Satisfaction",cex.aixs=4,cex.lab=4)+
   theme_classic()

#----- Box plot with error bar
  ggplot(data, aes(x=treetment, y=satisfaction)) +
   stat_boxplot(geom = "errorbar",width = 0.15) +
   labs(title="",x="Treatment", y = "Satisfaction",cex.aixs=4,cex.lab=4)+
   geom_boxplot()

#----- violin plot
  ggplot(data, aes(x=treetment, y=satisfaction)) +
   geom_violin(trim=FALSE, fill='#A4A4A4', color="darkred")+
   geom_boxplot(width=0.1) +
   labs(x="Treatment", y = "Satisfaction",cex.aixs=4,cex.lab=4)+
   theme_classic()
```


[DOI: 10.18502/jsbch.v9i1.18706]