© IJARW | ISSN (O) - 2582-1008
March 2026 | Vol. 7 Issue. 9
www.ijarw.com

USING GEOGEBRA TO CONSTRUCT CHARTS AND ANALYZE STATISTICAL DATA IN ORDER TO VISUALIZE DATA IN TEACHING DESCRIPTIVE STATISTICS IN PROBABILITY AND STATISTICS AT THE UNIVERSITY

Vi Dieu Minh

Faculty of Interdisciplinary Sciences, Thai Nguyen University of Agriculture and Forestry, Vietnam

ABSTRACT

In this paper, the author focuses on presenting three of the most commonly used charts in statistical data analysis: the histogram, the box plot, and the scatter plot. The paper not only introduces the significance of these three types of charts but also provides guidance on how to construct them using the dynamic geometry software GeoGebra, as well as how to interpret and extract information from them.

Keyword: Histogram, Box plot, Scatter plot, Geogebra, analyze statistical data.

1. INTRODUCTION

The field of Statistics constantly faces challenges posed by issues in science and industry. Challenges in storing, organizing, and retrieving data have led to the emergence of a new field called **'data mining.'** Massive amounts of data are being generated across various sectors, and the task of statisticians is to make them meaningful: extracting key patterns and trends while understanding **'what the data is telling us.'**[1]

As data plays an increasingly vital role across numerous fields, statistical data analysis has become an indispensable tool for making accurate and scientific decisions. Through the analytical process, learners can identify distribution patterns, trends, and latent relationships within the data. In particular, the use of charts helps visualize information clearly, making data comprehension and interpretation easier and more effective. Consequently, charts are not merely illustrative tools but also essential means of exploring and communicating the significance of statistical data.

Currently, numerous software applications are utilized for statistical data analysis and charting, such as Excel, SPSS, Minitab, R, and GeoGebra. Among these, GeoGebra is a free mathematics software developed to support teaching and learning in a more visual and dynamic manner. Furthermore, this software enables teachers to illustrate lessons more vividly and intuitively for students. GeoGebra is suitable for various educational levels, from primary school to university, and is widely used worldwide due to its convenience, ease of use, and free access. In this article, the author focuses on guiding the construction and interpretation of several characteristic charts in statistical data analysis—such as histograms, box plots, and scatter plots—while providing illustrative examples to help readers easily visualize and better understand the significance of these chart types.

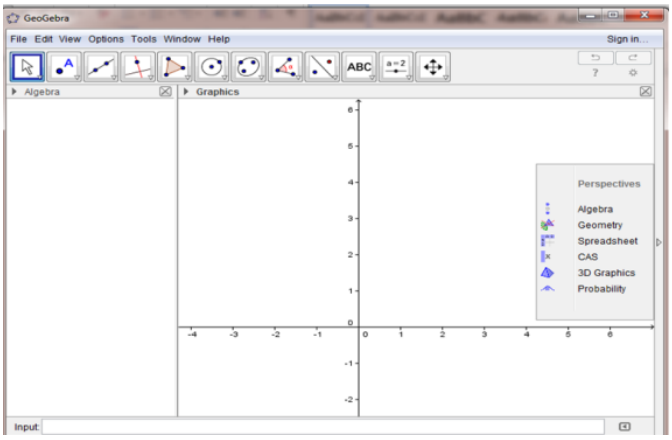
2. CONTENT

2.1 Installation Guide for GeoGebra and Its Significance in Mathematics Education.

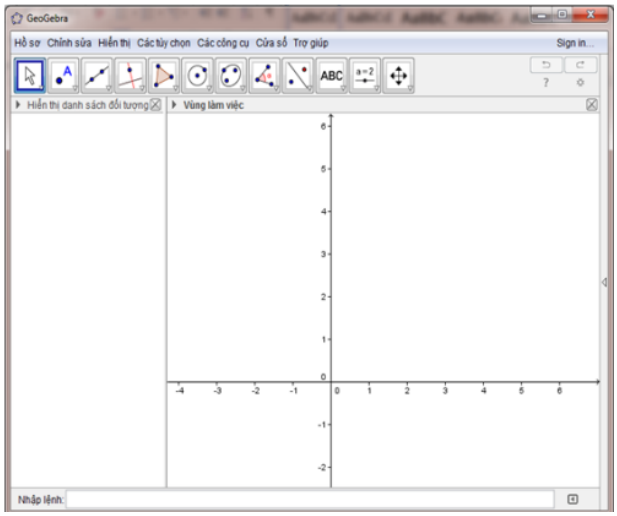
GeoGebra is a dynamic mathematics software designed to support the teaching and learning of mathematics from primary school to university. It seamlessly integrates **Geometry, Algebra, Calculus, and spreadsheets**. With its powerful yet user-friendly features, GeoGebra combines a computer algebra system (CAS), interactive geometry tools, and spreadsheets, allowing users to save time and computer storage space. Notably, users can create custom tools tailored to their specific needs. Furthermore, GeoGebra boasts a vast community with a rich repository of resources shared by users worldwide, facilitating the exchange of ideas and making mathematics education more accessible and effective.[2]

Installation and using guide for GeoGebra

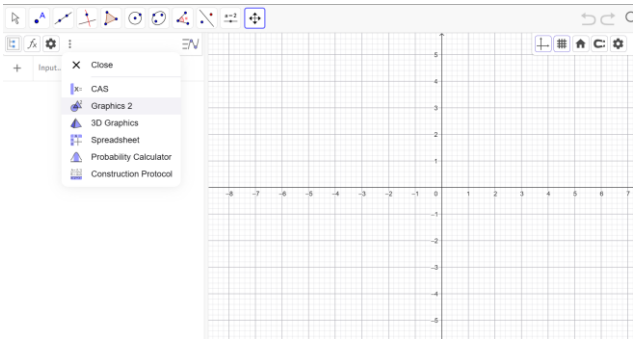
Step 1: Go to http://www.geogebra.org/download to download the software to your computer. After the installation is complete, select "Run"; GeoGebra will then launch and display the interface as shown below.



Step 2: To change the language (for example, from English to Vietnamese): click on "Options" in the menu bar, select "Language," then choose "R-Z" and click on "Vietnamese/Tiếng Việt." The interface will switch to Vietnamese, as shown below.



Step 3: Select the workspace: When the program launches, a Perspectives panel will appear, allowing you to choose your workspace, such as Algebra & Graphics, Geometry, 3D Graphics, Probability & Statistics, and more. You can show or hide the Perspectives panel by clicking the arrow icon on the right edge of the window to switch to a different workspace. In the Algebra & Graphics mode, the Input Bar is located at the bottom of the window, used for entering direct commands for drawing or calculations (as shown below).



2.2 Common Chart Types in Statistical Data Analysis.

Charts in statistical data analysis play a vital role in transforming complex raw data into visual representations, making it easier to identify trends, compare values, and quickly detect correlations or data anomalies. They simplify information, supporting more accurate and effective decision-making. In this article, the author focuses exclusively on three types of charts: **Histograms**, **Box plots**, and **Scatter plots**.

2.2.1 Histogram

A **histogram** is a type of chart that represents the distribution of numerical variable values as a series of bars. Each bar typically covers a range of numerical values known as a **bin** or **class**; the height of the bar indicates the frequency of data points within that corresponding interval. Histograms provide deep insights into the underlying patterns, trends, and characteristics of a dataset. They are particularly useful for understanding the frequency or volume of data across different intervals or groups.

a. How to Construct a Histogram in GeoGebra

- For Discrete Data: To create a frequency chart in GeoGebra, follow these steps:

1. **Open Spreadsheet:** Launch GeoGebra, go to the **View** menu, and select **Spreadsheet**.
2. **Input Data:** Enter the data values into **Column A** and their corresponding frequencies into **Column B**.
3. **Select Data:** Highlight (select) both columns of the entered data.
4. **One Variable Analysis:** Click the **One Variable Analysis** icon (the bar chart icon) on the spreadsheet toolbar.

5. **Create Chart:** In the new window that appears, select **"Bar Chart"** from the drop-down menu to display the frequency distribution.

- For Continuous Data (Grouped Data): To construct a histogram in GeoGebra for grouped data, follow these steps:

1. **Open Spreadsheet:** Launch GeoGebra, go to the **View** menu, and select **Spreadsheet**.
2. **Input Data (Grouped Data):**
    - **Column A (Class Boundaries):** Enter the boundaries (the lower limit of the first group and the upper limits of all subsequent groups).
    - **Column B (Frequencies):** Enter the corresponding frequencies for each group. (*Note: Column B will have one fewer value than Column A*).
3. **Create Data Lists:** Highlight the cells in each column separately, right-click, and select Create -> List. (Repeat this for both the boundaries and the frequencies to create list1 and list2).
4. **Construct the Histogram:** In the **Input Bar**, type the following command:

Histogram[<List of Class Boundaries>, <List of Frequencies>]

b. Interpreting Histograms

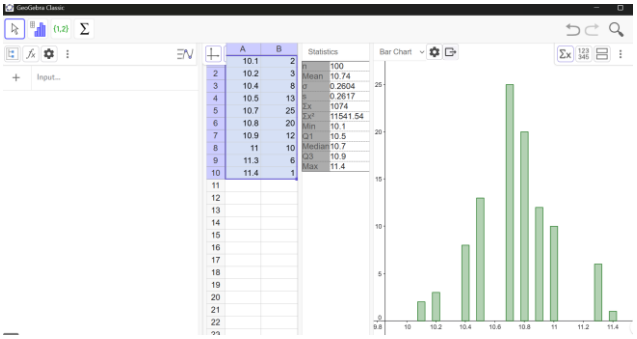
Analyzing a histogram opens the door to profound data insights. The **shape**—whether symmetrical or skewed—suggests underlying data patterns. The **peaks** (modes) mark the most frequent values, while the **width** represents the spread. **Outliers** stand out as isolated bars. The height of each bar reflects data frequency, and the horizontal axis defines the range of values. The **bin width** directly affects the level of detail. From a histogram, we can analyze several key characteristics:

- **Data Distribution:** Data distribution refers to how data points are spread across different values or ranges. By visualizing the distribution, you can grasp the frequency of specific values, the overall range, and the general behavior of the data.
- **Identifying Central Tendency and Dispersion: Central tendency** highlights the central or typical value around which data points tend to cluster. Common measures include the mean, median, and mode. On the other hand, **dispersion** (or variability) shows how data points are scattered around that central value.
- **Identifying Skewness and Symmetry:**
    - **Normal Distribution:** Often represented by a **bell-shaped curve** in a histogram, this distribution shows data that is symmetrical around a central point, with most values concentrated near the mean.
    - **Skewness:** This becomes evident in histograms with unequal tails.
        - **Positive Skew (Right-skewed):** The tail extends to the right, while the majority of data points are concentrated on the left. This indicates the dataset contains a few unusually high values that pull the mean toward the right.
        - **Negative Skew (Left-skewed):** The distribution leans toward higher values, with the tail extending to the left. This suggests the presence of a few extremely low values, pulling the mean toward the left.

Example 1: A survey of the monthly turnover X (in millions of VND) of 100 business households yielded the following data table:[4]

X	10,1	10,2	10,4	10,5	10,7	10,8	10,9	11	11,3	11,4
ni	2	3	8	13	25	20	12	10	6	1

To construct a histogram for discrete data, follow the steps outlined in section (a); the obtained results are as follows:

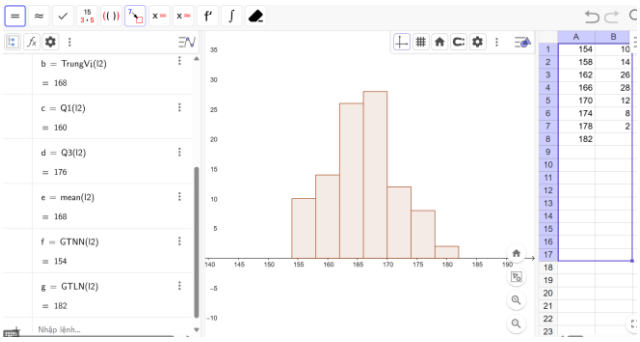


Based on the chart above, we can draw several observations: The dataset has a mean ≈ 10.74 and a median ≈ 10.7, indicating a relatively symmetrical distribution. Most values are concentrated within the [10.5; 10.9] range, with low dispersion (σ ≈ 0.26). The chart follows a near bell-shaped curve, slightly skewed to the right (as there are still values in the 11.2–11.4 range), and contains no outliers.

Example 2: A measurement of the height X (cm) of 100 young adults aged 18 to 22 in Province A yielded the following data table:[4]

X (cm)	154 -158	158-162	162-166	166-170	170-174	175-178	178-182
ni	10	14	26	28	12	8	2

To construct a histogram for continuous data, follow the steps outlined in section (a); the obtained results are as follows:



Based on the histogram above, we can draw several observations: The dataset has a mean ≈ median ≈ 168, indicating a relatively symmetrical distribution. Most values are concentrated within the 160–176 range, with a low degree of dispersion. The chart follows a near bell-shaped curve, where frequencies peak around 165–170 and then gradually decrease toward both ends, showing that the data is strongly clustered around the central value. No outliers are observed.

2.2.2 Box plots

A Box plot, also known as a 'Box and whisker plot,' is a chart that represents five key positions of a data distribution: the minimum value (min), the first quartile (Q1), the median, the third quartile (Q3), and the maximum value (max). The distance between Q1 and Q3 (known as the Interquartile Range – IQR) reflects the variability of the majority of the data, while points lying outside the interval Q1 – 1.5×IQR or Q3 + 1.5×IQR are typically considered outliers. Consequently, box plots allow for the rapid detection of uneven dispersion, distribution skewness, and anomalies.

a. How to Construct a Box Plot in GeoGebra

To create a box plot in GeoGebra, follow these steps:

1. **Open Spreadsheet:** Launch GeoGebra, go to the **View** menu, and select **Spreadsheet**.
2. **Input Data:** Enter the data values into **Column A** and their corresponding frequencies into **Column B**.
3. **Select Data:** Highlight (select) both columns of the entered data.
4. **One Variable Analysis:** Click the **One Variable Analysis** icon (the bar chart icon) on the spreadsheet toolbar.
5. **Create Chart:** In the new window that appears, select **"Box Plot"** from the drop-down menu to display the box plot.

b. Interpreting a Box Plot

- **The Median:** The line inside the box represents the **median** of the data. Half of the data points lie above this value, and half lie below. If the data is symmetrical, the median will be centered within the

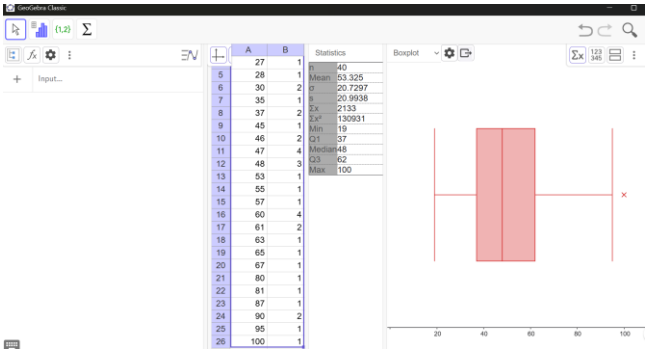
box. If the data is skewed, the median will appear closer to either the top or the bottom of the box.

- **The Box (Quartiles):** The bottom and top of the box represent the **25th and 75th percentiles**, respectively. These are also known as **quartiles** because each cuts off one-quarter (25%) of the data. The length of the box is the difference between these two percentiles, known as the **Interquartile Range (IQR)**.
- **The Whiskers:** The lines extending from the box are called **whiskers**. They represent the expected variability of the data. Whiskers extend up to $1.5 \times \text{IQR}$ from the top and bottom of the box. If the data does not reach the full length of the whiskers, they will simply end at the minimum and maximum data values.
- **Outliers:** Any data points located above or below the ends of the whiskers are plotted as individual **"x"** marks. These are called **outliers**—extreme values that fall outside the expected variation. These points should be carefully reviewed to determine if they are genuine anomalies or data entry errors; the whiskers do not include these outliers.

Example 3: The following dataset represents the test scores of a group of students (on a 100 -point scale):[5]

61	27	26	37	30	47	87	90
63	46	67	19	81	47	100	25
45	60	65	53	35	28	80	95
57	37	45	25	48	60	48	47
30	47	60	61	55	48	60	90

To construct a box plot for this data, follow the steps outlined in section (a); the obtained results are as follows:



Based on the box plot above, we can draw the following observations: The median ≈ 48, Q1 ≈ 37, Q3 ≈ 62, and the Interquartile Range (IQR) = Q3 − Q1 ≈ 25. The data shows a relatively wide dispersion (ranging from 19 to 100), with 50% of the values concentrated between 37 and 62. The distribution exhibits a slight right-skewed tendency and shows a significant outlier on the right side

2.2.3 Scatter plot

A Scatter plot is a collection of points M(x; y) in a rectangular coordinate system. It is a visual tool that uses the Cartesian coordinate system to display the correlation between two quantitative variables. Each data point on the chart represents a pair of values (x; y), helping to identify trends, the degree of correlation (positive/negative), and detect outliers.

Based on the scatter plot, we can determine the type of relationship between two quantities X and Y. If the points M(x; y) cluster around a straight line d, we say that the two random variables X and Y have a linear correlation. In this case, line d is referred to as the linear regression line.

a. How to Construct a Scatter Plot in GeoGebra

To create a scatter plot and find the regression line in GeoGebra, follow these steps:

1. **Open Spreadsheet:** Launch GeoGebra, go to the **View** menu, and select **Spreadsheet**.
2. **Input Data:** Enter the x values into **Column A** and the corresponding y values into **Column B**.
3. **Select Data:** Highlight (select) both columns of the entered data.
4. **Two Variable Regression Analysis:** Click the **Data Analysis** icon (bar chart icon) on

the spreadsheet toolbar and select **Two Variable Regression Analysis**.

5. **Construct the Linear Regression Line:** Within the regression analysis window, choose "Linear" from the **Regression Model** drop-down menu, or use the input command:

FitLine(<List of Points>) to draw the regression line through the data points.

b. Interpreting a Scatter Plot

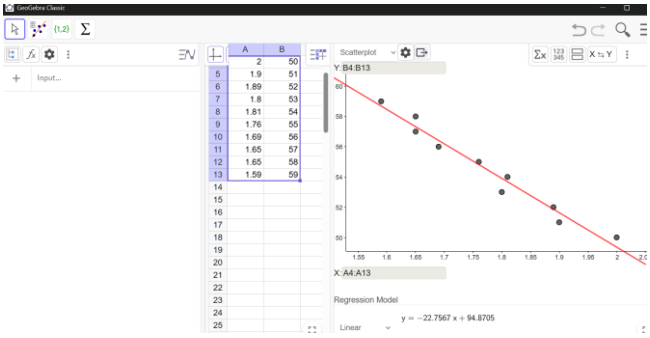
A **Scatter Plot** focuses on identifying the relationship between two variables by examining the direction, form, and strength of the data points.

- **Direction:**
  - **Positive Correlation:** Data points move upward from left to right (as X increases, Y increases).
  - **Negative Correlation:** Data points move downward from left to right (as X increases, Y decreases).
  - **No Correlation:** Data points are scattered randomly with no discernible pattern.
- **Form:**
  - **Linear:** The points cluster along a straight line.
  - **Non-linear (Curvilinear):** The points form a curved pattern (e.g., exponential or quadratic).
- **Strength:**
  - **Strong:** Data points lie very close to each other or form a clear, tight line.
  - **Weak:** Data points are widely scattered far from the trend line.
  - **Correlation Coefficient (r):** An r value close to -1 or 1 indicates a strong relationship; an r value close to 0 indicates a weak relationship.
- **Outliers:** Look for points located far away from the main cluster; these are special cases that require attention.
- **Trend Line (Line of Best Fit):** Drawing a line through the middle of the data points helps to more clearly define the overall trend of the relationship.

Example 4: In a specific crossbred pig breed, researchers are interested in two criteria: the lean meat percentage (%), denoted as Y, and the backfat thickness (cm), denoted as X. Measurements taken from 10 pigs yielded the following results:[6]

X (cm)	2	1,9	1,89	1,8	1,81	1,76	1,69	1,65	1,65	1,59
Y (%)	50	51	52	53	54	55	56	57	58	59

To construct a scatter plot for this data, follow the steps outlined in section (a); the obtained results are as follows:



Based on the scatter plot above, we can draw the following observations: The data points lie very close to a downward-sloping line, indicating a strong negative linear correlation between the two variables. As X increases (from approximately 1.59 → 2), Y decreases (from approximately 59 → 50). From the regression equation y = -22.76x + 94.8705, it is evident that for every 1-unit increase in X, Y decreases by an average of approximately 22.76 units. The fact that the points do not deviate significantly from the line suggests that the regression model fits the data very well (low noise). In conclusion, the two variables share

a strong inverse relationship, and the linear model can be used for highly accurate predictions.

3. CONCLUSION

Visual tools such as histograms, box plots, and scatter plots are essential for effectively visualizing and analyzing data. They enable the identification of distributions, variability, outliers, and relationships between variables. The integration of GeoGebra in constructing these charts enhances both visual clarity and interactivity within the classroom. Consequently, learners can more easily access knowledge while developing analytical thinking and data processing skills. This approach significantly contributes to improving the quality of teaching and learning statistics in school

REFERENCES

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

[2] Department of Mathematics – School of Information and Communication Technology. (2024). *GeoGebra Dynamic Software*. Can Tho University. [Online]. Available: https://se.ctu.edu.vn/images/upload/bmtoan/phanmem/GEOGEBRA.pdf

[3] Ta, D. P. (2026). *Teaching and Learning Mathematics with GeoGebra*. Vietnam Education Publishing House.

[4] Nguyen, T. D., Mai, T. N. H., Vi, D. M., & Bui, L. P. (2023). *Probability and Statistics Exercises and Practice on R Software*. Hanoi University of Science and Technology (HUST) Publishing House.

[5] Phan, T. H., & Nguyen, T. N. (2018). *Applied Statistics: A Practical Guide Using R Software*. Statistics Publishing House.

[6] Nguyen, T. D., Mai, T. N. H., & Pham, T. H. (2018). *Probability and Statistics*. Hanoi University of Science and Technology (HUST) Publishing House.

[7] Department of Mathematics and Physics – Thai Nguyen University of Agriculture and Forestry. (2018). *Advanced Mathematics Lecture Notes*. (Internal Circulation Document).

[8] Le, D. T. (2005). *Advanced Mathematics for Economists*. Statistics Publishing House.