Learning Stata: A Tutorial on Creating and Customizing Histograms for Data Visualization

Authored by

Mohammed looti

November 8, 2025

A **histogram** is an indispensable graphical tool within statistical analysis, serving as the foundational method for visualizing the empirical distribution of a continuous dataset. Through the use of connected rectangular bars, this chart effectively depicts the frequency, count, or proportion of data observations that fall within specific, predetermined numerical intervals, commonly referred to as bins. The insights gained from examining a histogram--specifically the data's shape (symmetry or skewness), center (mean or median location), and spread (variance)--are crucial for hypothesis testing, outlier detection, and subsequent accurate statistical modeling. For any exploratory data analysis (EDA), generating a well-constructed histogram is often the essential first step.

This comprehensive tutorial is specifically tailored for quantitative analysts, researchers, and students who utilize **Stata**, one of the most powerful and widely respected statistical software packages in academia and professional research. We will guide you through the initial steps of basic histogram creation and then introduce a suite of powerful commands designed to modify the visualization's scale, refine its granularity, and apply essential statistical overlays. Mastery of these techniques ensures that your data visualizations move beyond simple plots to become highly informative and professionally presentable assets.

Preparing the Environment and Loading Sample Data in Stata

Before we embark on creating visualizations, it is standard practice to set up the statistical environment and gain familiarity with the dataset slated for analysis. For the purpose of maintaining consistency and ensuring reproducibility throughout this tutorial, we will rely on the standard, internal dataset provided within the Stata software suite, known as *auto*. This dataset contains comprehensive specifications for 74 distinct automobiles and is frequently used in Stata documentation for illustrative examples, making it instantly accessible to all users.

To load this illustrative dataset directly into your current Stata session, execute the following command precisely as shown in the Command window:

use http://www.stata-press.com/data/r13/auto

Once the data has been successfully loaded, the next vital step involves conducting a quick structural and statistical overview. The robust **summarize** command is instrumental here, providing key descriptive statistics for every variable contained within the active dataset. This output includes essential metrics such as the total number of observations, the arithmetic mean, the standard deviation (a measure of spread), and the overall range. Executing this step confirms data integrity and helps the analyst quickly pinpoint the continuous variables best suited for visualization via histogram.

To review the dataset overview and confirm its readiness, enter the following command:

summarize

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)

. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

As clearly illustrated in the output generated by the **summarize** command, the *auto* dataset comprises 12 variables in total, which encompass a mix of numerical attributes (like vehicle length and weight) and categorical indicators. For the remainder of this guide, we will specifically concentrate on the continuous variable labeled *length* to effectively demonstrate the creation and customization process for histograms in Stata.

Generating Core Histograms: Syntax and Default Settings

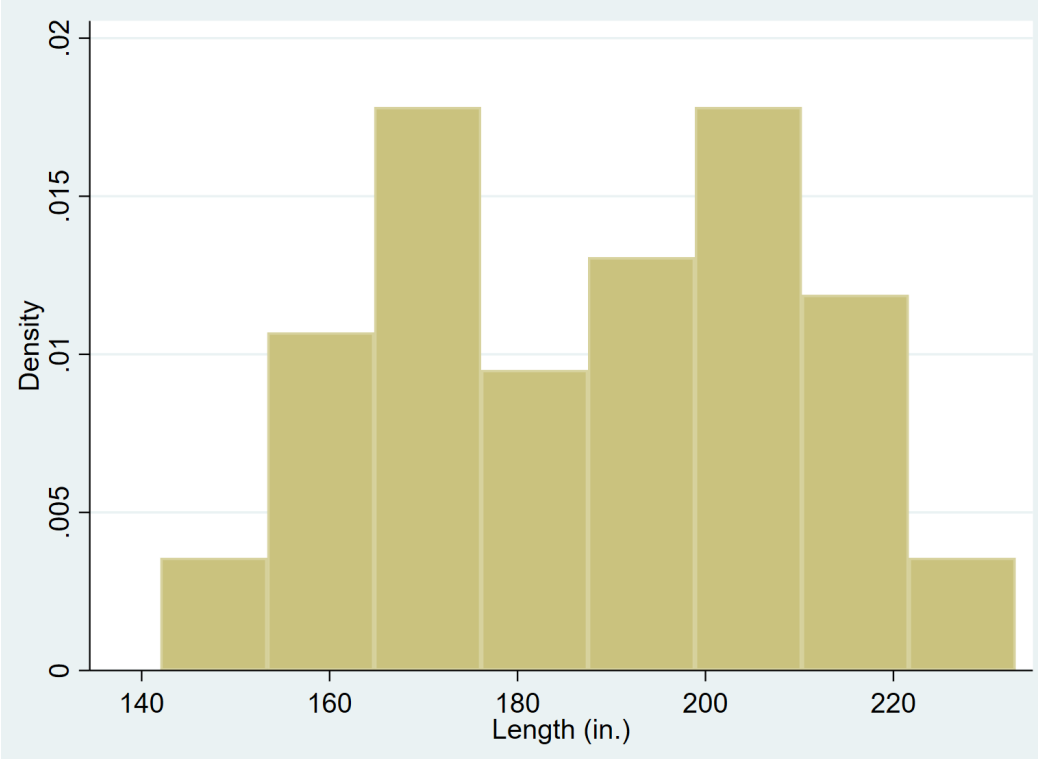
The foundation of histogram generation in Stata rests upon the straightforward **hist** command. To execute the command, the analyst needs only to specify the name of the variable they wish to plot. By default, Stata employs sophisticated internal algorithms to make several critical decisions regarding the graph's initial presentation. These decisions include calculating an optimal number of bins (intervals) for the data aggregation and setting the scaling of the vertical (Y) axis. The typical default scaling for the Y-axis is density, which is highly useful when comparing the shapes of distributions across different sample sizes.

Generating the Default Histogram

To create the most basic visualization depicting the distribution of vehicle lengths, we simply pair the base command with the target variable name. This initial, unedited plot serves as a rapid visual assessment tool, allowing researchers to quickly identify major characteristics of the data's shape, such as pronounced skewness, potential modality issues, or the presence of significant outliers.

Execute the following command to plot the default visualization for the *length* variable:

hist length



Controlling the Vertical Axis: Frequencies and Percentages

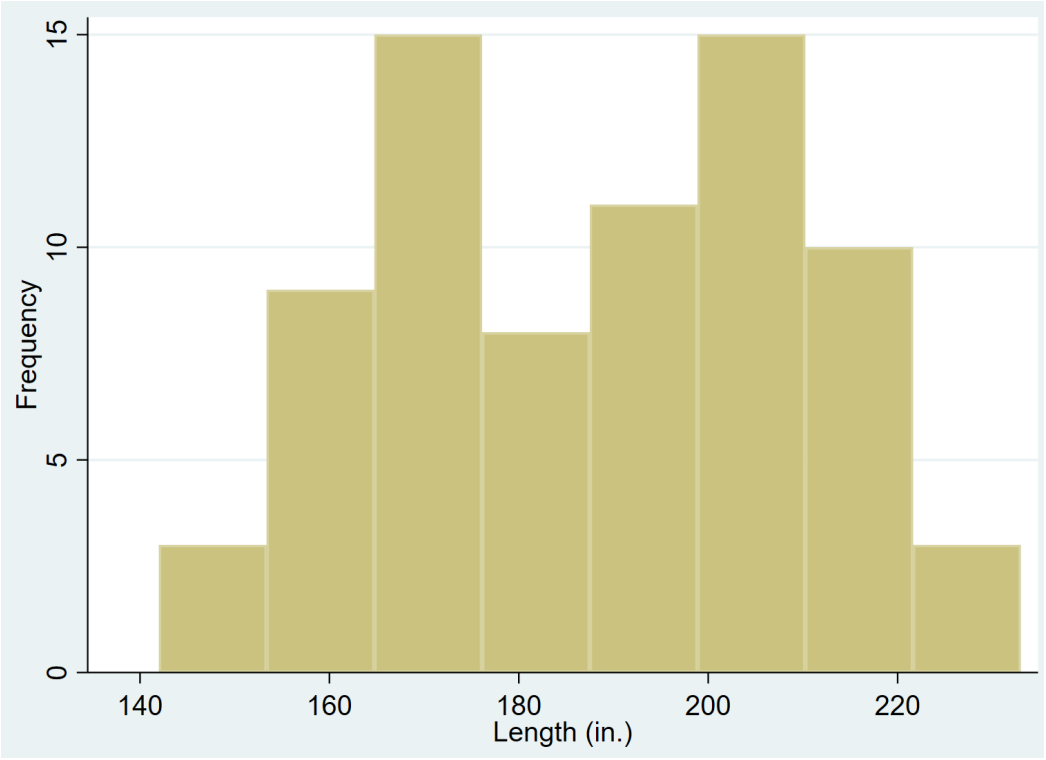
Although the default histogram utilizes density on the Y-axis, which is statistically appropriate for probability distributions, analysts often require different metrics depending on the analytical goal. Stata provides easy-to-use options to switch the vertical scale to display raw counts or relative proportions, significantly altering the graph's interpretation.

Adjusting the Y-Axis to Display Frequencies

For scenarios where the exact number of observations falling into each interval is the primary interest, researchers prefer to display absolute counts, often referred to as frequencies. This modification is easily achieved in Stata by applying the **freq** option. Switching to frequencies is particularly valuable when the analysis focuses on the numerical magnitude of occurrences within specific data ranges, rather than the probabilistic density.

To instruct Stata to display the absolute frequencies, append the **freq** option to the **hist** command, separated by a comma:

hist length, freq

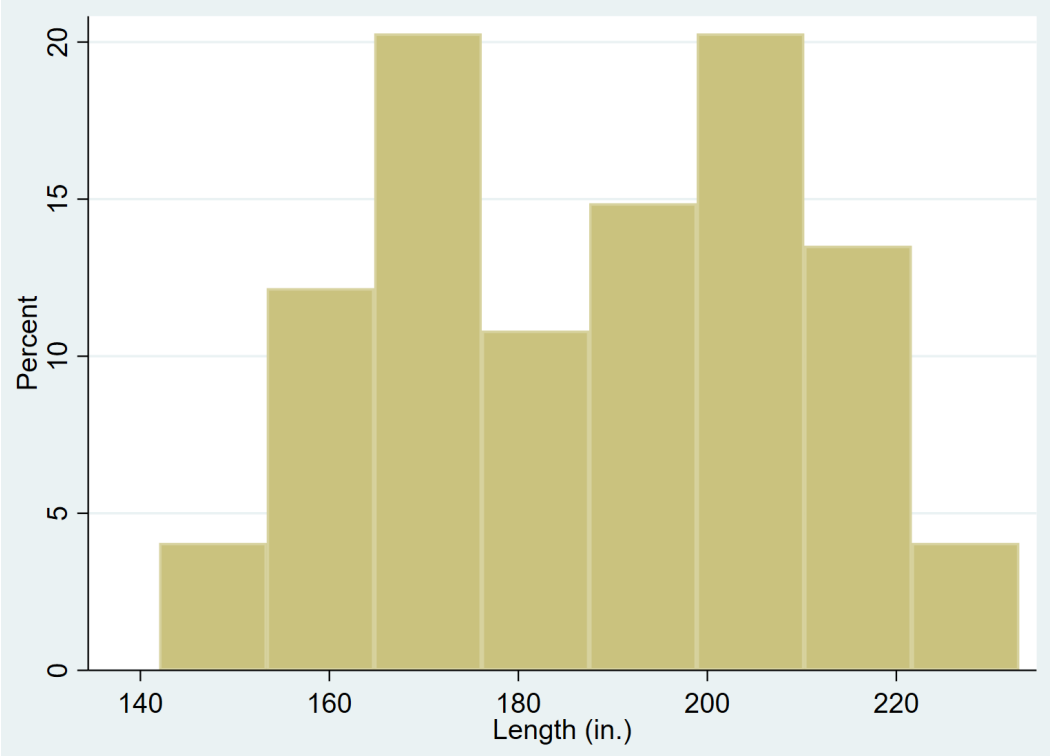


Using Percentages for Relative Comparison

An alternative and highly effective method for conveying distributional characteristics, especially to a non-technical audience, is scaling the Y-axis to display percentages. When the goal is to understand the proportion of the total sample contained within each bin, the **percent** option converts the raw counts into relative frequencies. This provides an immediate sense of the data's composition relative to the entire dataset and is ideal for comparing how different subgroups are distributed.

To switch the vertical axis to display percentages, utilize the **percent** option:

hist length, percent



Refining Granularity: Mastering Bin Control with the bin() Option

The crucial factor dictating the visual appearance and, critically, the interpretability of a histogram is the selection of the number of bins--the intervals used to group the continuous data. By default, Stata applies established statistical rules, such as the Sturges formula, to automatically calculate an optimal number of bins (often 8, as seen in the earlier examples) that should theoretically provide a balanced view of the underlying data distribution:

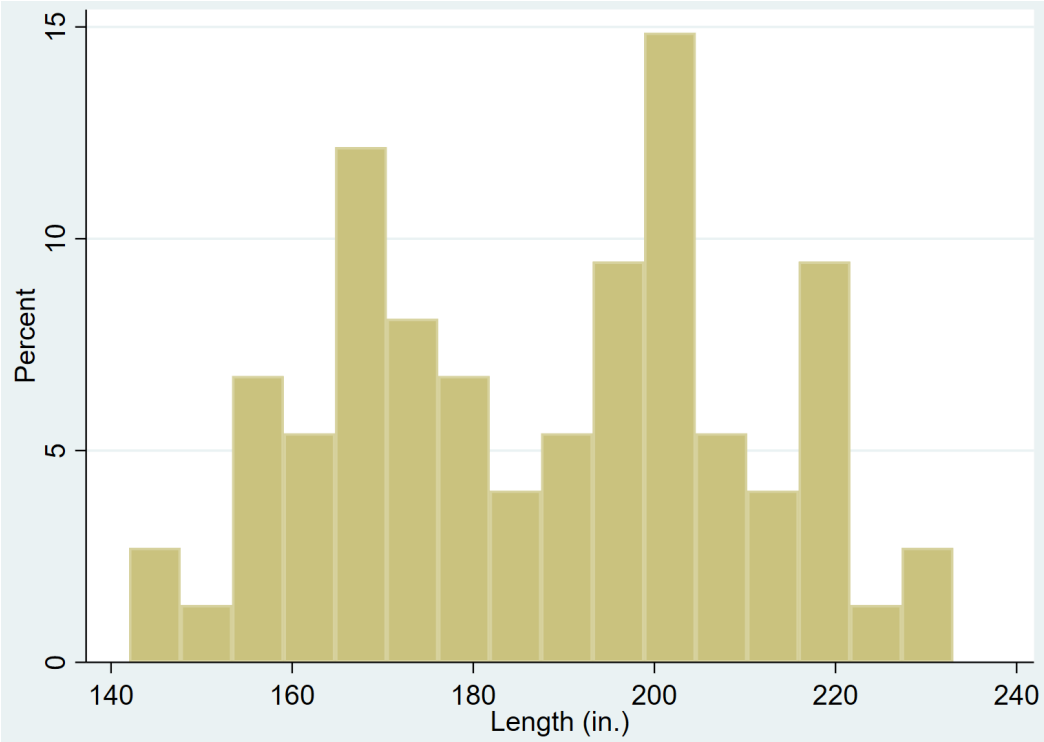
```
. hist length
(bin=8, start=142, width=11.375)

. hist length, freq
(bin=8, start=142, width=11.375)

. hist length, percent
(bin=8, start=142, width=11.375)
```

However, analysts often need to override this default setting to achieve specific visual goals. Using too few bins can lead to over-smoothing, potentially obscuring important underlying features of the

data's shape. Conversely, using an excessive number of bins can make the visualization overly jagged, sensitive to minor fluctuations or noise, and difficult to interpret. The **bin()** option grants the user precise control over this key parameter, enabling them to customize the level of detail presented.

To increase the level of detail and granularity in the plot, we can specify a larger number of bins. The following command instructs Stata to use 16 bins, doubling the default count, which provides a much more granular look at the data's dispersion across the entire range of vehicle lengths:

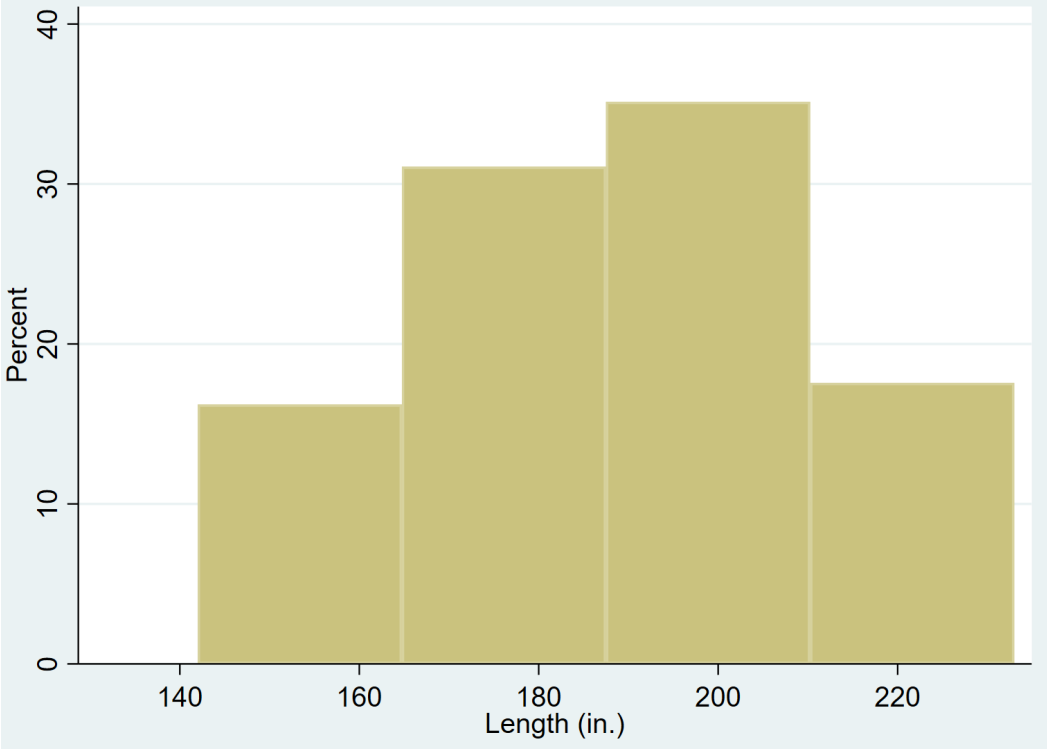
hist length, percent bin(16)



Conversely, if the objective is to simplify the histogram and provide a smoother, broader view that emphasizes the data's central tendency without distraction from fine detail, we should decrease the number of bins. This approach aggregates observations into wider intervals, effectively highlighting only the major peaks of the distribution.

Here is the command to reduce the bin count significantly to four:

hist length, percent bin(4)



It is essential for the analyst to recognize the inherent trade-off in bin selection: fewer bins result in a smoother, more general outline, while a greater number of bins provides higher granularity, allowing for the visualization of minor fluctuations in the data's frequency. Choosing the correct number of bins is an interpretive and iterative process, guided primarily by the specific analytical question being addressed.

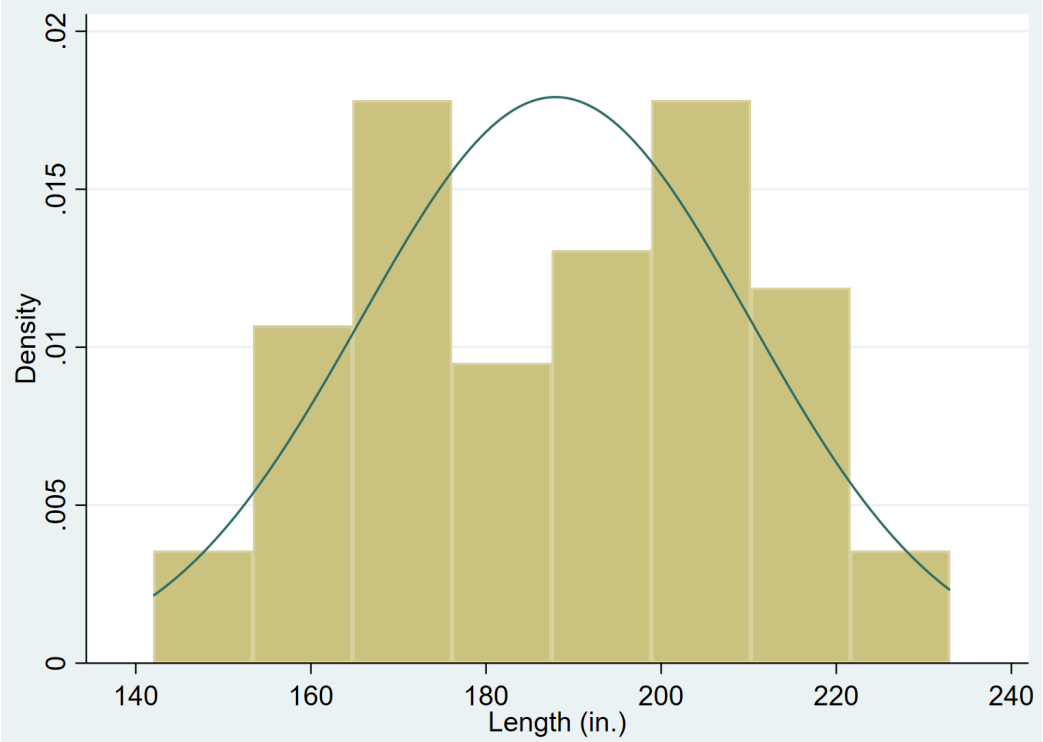
Adding Statistical Overlays: Assessing Normality with the Density Curve

A primary concern in many statistical applications is determining whether the empirical data distribution closely approximates a theoretical model, most notably the normal density curve (often referred to as the Gaussian distribution). Since the assumption of normality underlies a vast number of parametric statistical tests, visually comparing the observed data to this theoretical curve is highly informative and often mandatory.

Stata makes this visual comparison straightforward by allowing the user to overlay a normal distribution curve directly onto the histogram. This curve is not generic; it is precisely fitted to the data using the calculated sample mean and standard deviation of the variable currently being plotted. The inclusion of this overlay is activated by simply adding the **normal** option to the **hist** command. It is important to note a critical functional requirement: when the **normal** option is utilized, Stata automatically overrides any previous scaling choices and reverts the y-axis scaling back to density. This is necessary to correctly plot the probability density function (PDF) of the theoretical curve alongside the empirical data.

Execute the following command to add the normal density overlay to the histogram of vehicle lengths:

hist length, normal



Achieving Publication Quality: Aesthetic Enhancements and Annotations

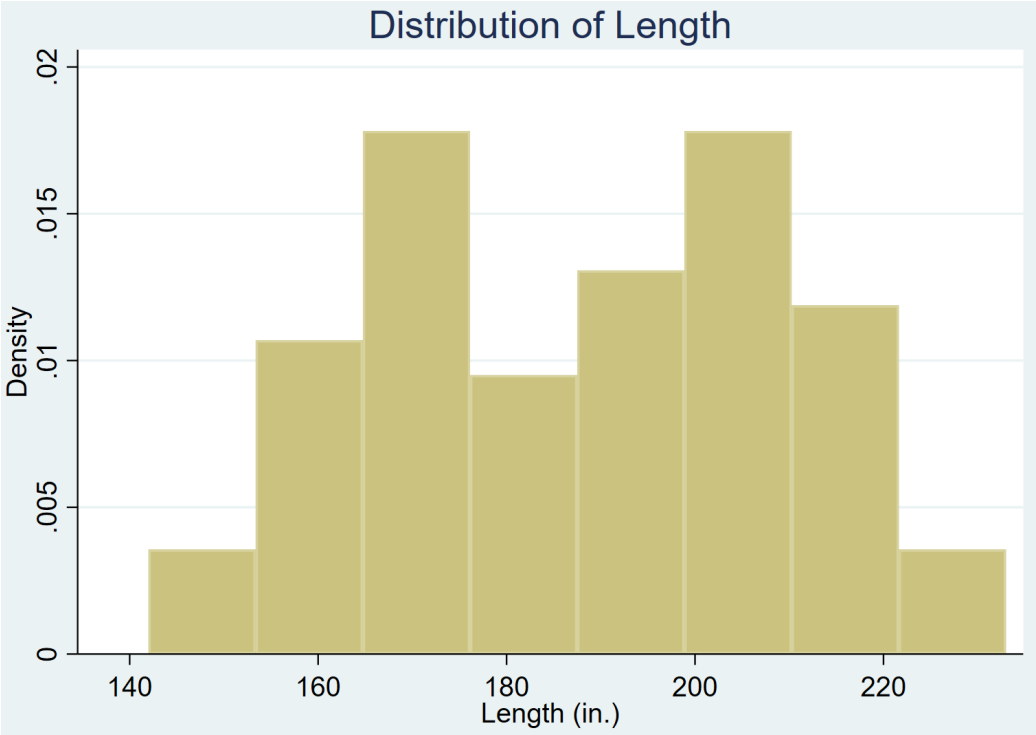
Beyond the core statistical visualization tools, Stata offers extensive options for enhancing the aesthetic appeal and clarity of graph elements. For any visual destined for professional reporting or publication, clear and comprehensive labeling is absolutely critical. We can significantly improve the appearance and communicative effectiveness of our histograms by incorporating informative titles, contextual subtitles, and proper explanatory notes using Stata's powerful built-in graph options.

Applying a Descriptive Title

A well-crafted title serves as the viewer's immediate guide, summarizing the content of the graph. The **title()** option allows you to place a descriptive string centered above the entire plot area. This is the primary labeling mechanism and should succinctly convey the main variable being analyzed and the nature of the measurement.

To add a title describing the distribution of vehicle lengths, use the following syntax:

hist length, title("Distribution of Length")

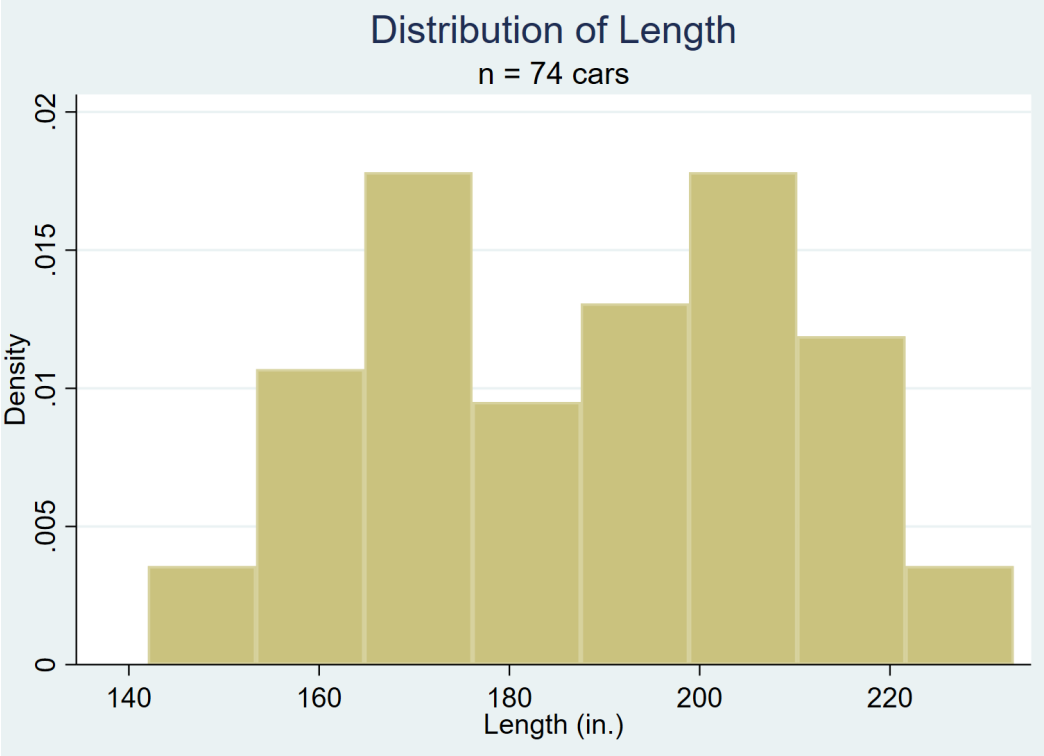


Including a Subtitle for Context

Subtitles are highly effective for providing secondary information that adds crucial context to the main title, such as defining the sample size, specifying the data collection period, or identifying the specific population under study. The **subtitle()** option positions this text directly beneath the main title, enabling the inclusion of greater detail without causing clutter in the primary label area.

We can efficiently add the sample size (n=74 cars) as a subtitle by including the **subtitle()** option alongside the main title command:

hist length, title("Distribution of Length") subtitle("n = 74 cars")

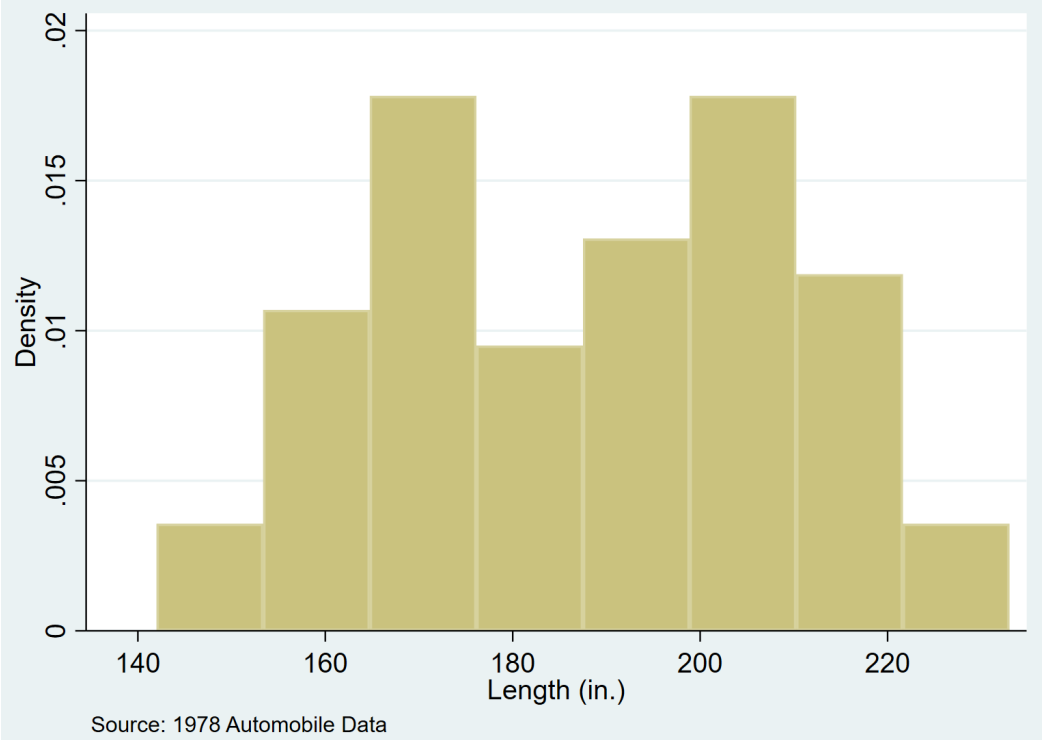


Adding Source Information via a Comment Note

Finally, adherence to professional standards dictates that all statistical graphs must include proper attribution and citation. The **note()** command allows the user to insert a small comment or footnote, typically placed at the bottom-left of the graph area. This is the ideal location to cite the original data source, offer a brief clarification on methodology, or acknowledge external contributors.

To include a note citing the origin of the *auto* dataset, use the **note()** command:

hist length, note("Source: 1978 Automobile Data")



Conclusion: Summarizing Stata Histogram Mastery

The ability to generate and customize histograms effectively is a core competency for any user of Stata involved in quantitative research. By starting with the simple **hist** command and progressively integrating options like **freq**, **percent**, **bin()**, and **normal**, researchers gain complete control over how their data's underlying distribution is visually represented. Furthermore, utilizing aesthetic options such as **title()**, **subtitle()**, and **note()** transforms raw statistical output into sophisticated, publication-ready graphics. These techniques ensure that your data storytelling is both statistically rigorous and visually compelling.