ENGINEERING STATISTICS HANDBOOK
HOME TOOLS & AIDS SEARCH BACK NEXT

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic

1.3.3.14. Histogram

Purpose: Summarize a Univariate Data Set

The purpose of a histogram (Chambers) is to graphically summarize the distribution of a univariate data set.

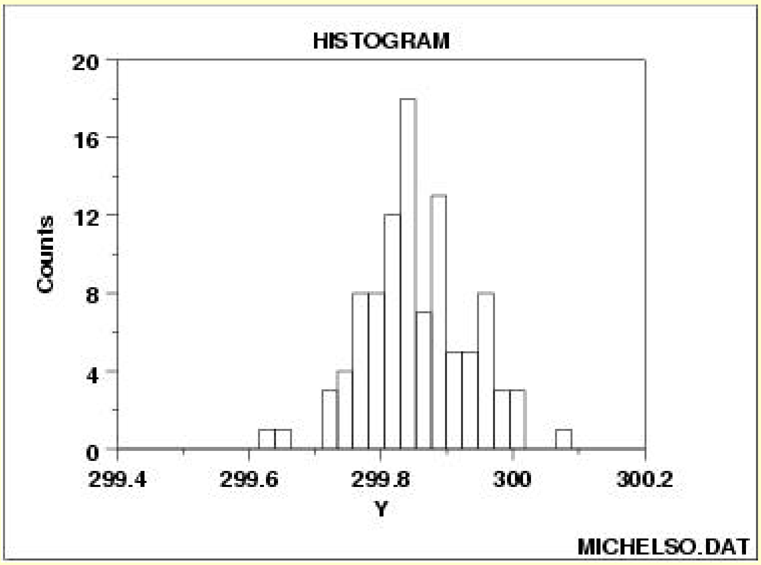
The histogram graphically shows the following:

1. center (i.e., the location) of the data;
2. spread (i.e., the scale) of the data;
3. skewness of the data;
4. presence of outliers; and
5. presence of multiple modes in the data.

These features provide strong indications of the proper distributional model for the data. The probability plot or a goodness-of-fit test can be used to verify the distributional model.

The examples section shows the appearance of a number of common features revealed by histograms.

Sample Plot



Definition

The most common form of the histogram is obtained by splitting the range of the data into equal-sized bins (called classes). Then for each bin, the number of points from the data set that fall into each bin are counted. That is

- Vertical axis: Frequency (i.e., counts for each bin)
- Horizontal axis: Response variable

The classes can either be defined arbitrarily by the user or via some systematic rule. A number of theoretically derived rules have been proposed by Scott (Scott 1992).

The cumulative histogram is a variation of the histogram in which the vertical axis gives not just the counts for a single bin, but rather gives the counts for that bin plus all bins for smaller values of the response variable.

Both the histogram and cumulative histogram have an additional variant whereby the counts are replaced by the normalized counts. The names for these variants are the relative histogram and the relative cumulative histogram.

There are two common ways to normalize the counts.

1. The normalized count is the count in a class divided by the total number of observations. In this case the relative counts are normalized to sum to one (or 100 if a percentage scale is used). This is the intuitive case where the height of the histogram bar represents the proportion of the data in each class.

2. The normalized count is the count in the class divided by the number of observations times the class width. For this normalization, the area (or integral) under the histogram is equal to one. From a probabilistic point of view, this normalization results in a relative histogram that is most akin to the probability density function and a relative cumulative histogram that is most akin to the cumulative distribution function. If you want to overlay a probability density or cumulative distribution function on top of the histogram, use this normalization. Although this normalization is less intuitive (relative frequencies greater than 1 are quite permissible), it is the appropriate normalization if you are using the histogram to model a probability density function.

Questions

The histogram can be used to answer the following questions:

1. What kind of population distribution do the data come from?

2. Where are the data located?
3. How spread out are the data?
4. Are the data symmetric or skewed?
5. Are there outliers in the data?

Examples

1. Normal
2. Symmetric, Non-Normal, Short-Tailed
3. Symmetric, Non-Normal, Long-Tailed
4. Symmetric and Bimodal
5. Bimodal Mixture of 2 Normals
6. Skewed (Non-Symmetric) Right
7. Skewed (Non-Symmetric) Left
8. Symmetric with Outlier

Related Techniques

Box plot
Probability plot

The techniques below are not discussed in the Handbook. However, they are similar in purpose to the histogram. Additional information on them is contained in the Chambers and Scott references.

Frequency Plot
Stem and Leaf Plot
Density Trace

Case Study

The histogram is demonstrated in the heat flow meter data case study.

Software

Histograms are available in most general purpose statistical software programs. They are also supported in most general purpose charting, spreadsheet, and business graphics programs. Dataplot supports histograms.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT