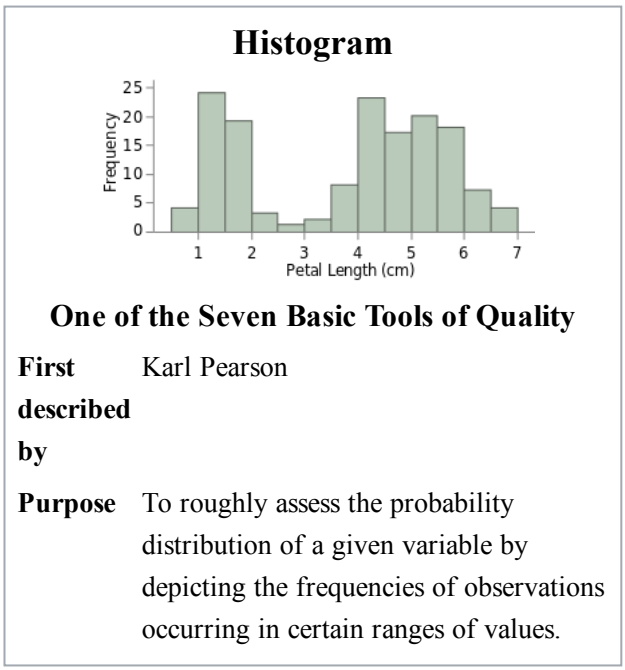
Histogram

From Wikipedia, the free encyclopedia

A **histogram** is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable) and was first introduced by Karl Pearson.[1] It is a kind of bar graph. To construct a histogram, the first step is to "bin" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent, and are often (but are not required to be) of equal size.[2]



If the bins are of equal size, a rectangle is erected over the bin with height proportional to the frequency — the number of cases in each bin. A histogram may also be normalized to display "relative" frequencies. It then shows the proportion of cases that fall into each of several categories, with the sum of the heights equaling 1.

However, bins need not be of equal width; in that case, the erected rectangle is defined to have its *area* proportional to the frequency of cases in the bin.[3] The vertical axis is then not the frequency but *frequency density* — the number of cases per unit of the variable on the horizontal axis. Examples of variable bin width are displayed on Census bureau data below.

As the adjacent bins leave no gaps, the rectangles of a histogram touch each other to indicate that the original variable is continuous.[4]

Histograms give a rough sense of the density of the underlying distribution of the data, and often for density estimation: estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the x-axis are all 1, then a histogram is identical to a relative frequency plot.

A histogram can be thought of as a simplistic kernel density estimation, which uses a kernel to smooth frequencies over the bins. This yields a smoother probability density function, which will in general more accurately reflect distribution of the underlying variable. The density estimate could be plotted as an alternative to the histogram, and is usually drawn as a curve rather than a set of boxes.

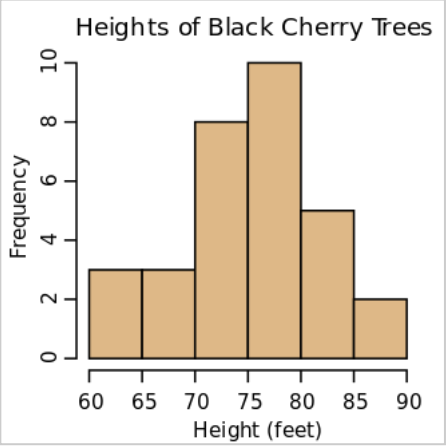
Another alternative is the average shifted histogram,[5] which is fast to compute and gives a smooth curve estimate of the density without using kernels.

The histogram is one of the seven basic tools of quality control.[6]

Histograms are sometimes confused with bar charts. A histogram is used for continuous data, where the bins represent ranges of data, while a bar chart is a plot of categorical variables. Some authors recommend that bar charts have gaps between the rectangles to clarify the distinction.

Contents

Etymology



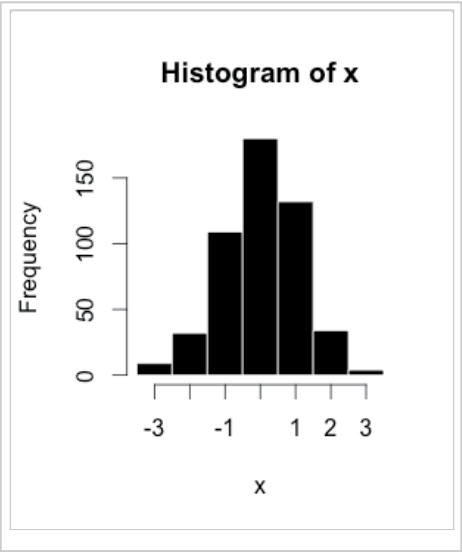
An example histogram of the heights of 31 Black Cherry trees.

The etymology of the word *histogram* is uncertain. Sometimes it is said to be derived from the Ancient Greek ἱστός (*histos*) – "anything set upright" (as the masts of a ship, the bar of a loom, or the vertical bars of a histogram); and γράμμα (*gramma*) – "drawing, record, writing". It is also said that Karl Pearson, who introduced the term in 1891, derived the name from "historical diagram".[7]

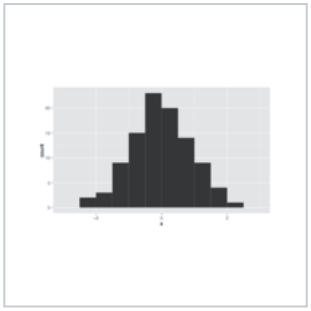
Examples

This is the data for the histogram to the right, using 509 items:

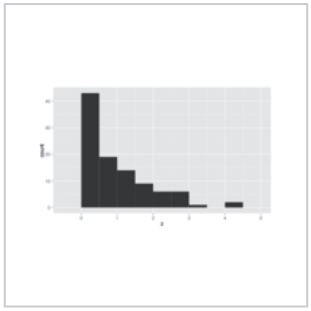
Bin	Count
−3.5 to -2.51	9
−2.5 to -1.51	32
−1.5 to -0.51	109
−0.5 to 0.49	180
0.5 to 1.49	132
1.5 to 2.49	34
2.5 to 3.49	4



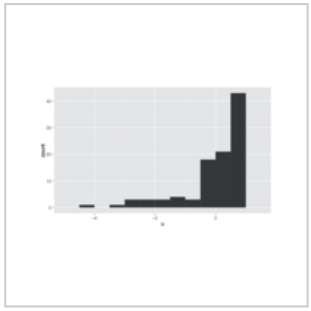
The words used to describe the patterns in a histogram are: "symmetric", "skewed left" or "right", "unimodal", "bimodal" or "multimodal".



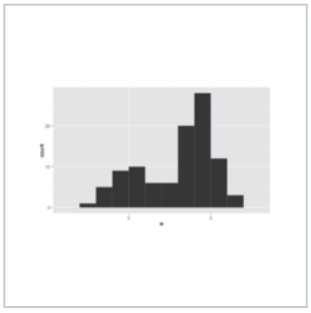
Symmetric, unimodal



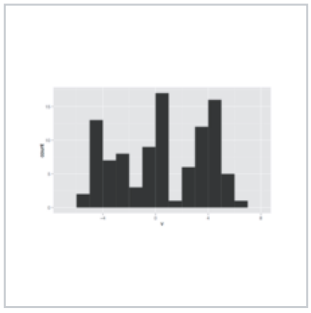
Skewed right



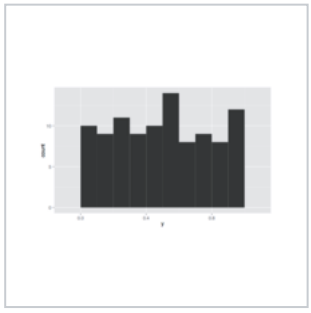
Skewed left



Bimodal

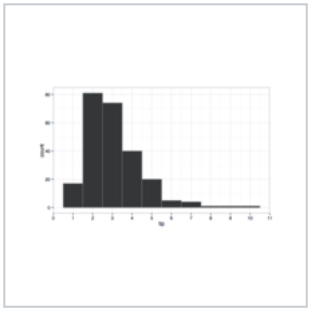


Multimodal

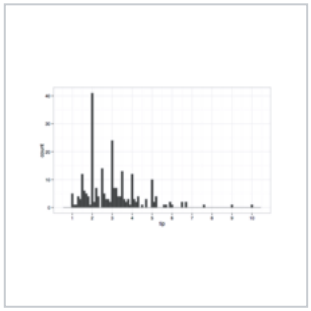


Symmetric

It is a good idea to plot the data using several different bin widths to learn more about it. Here is an example on tips given in a restaurant.

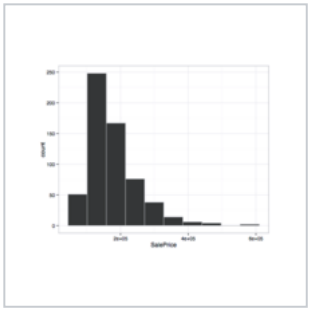


Tips using a $1 bin width, skewed right, unimodal

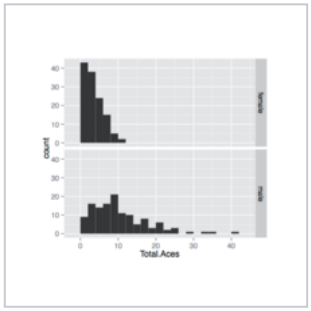


Tips using a 10c bin width, still skewed right, multimodal with modes at $ and 50c amounts, indicates rounding, also some outliers

Here are a couple more examples:



Prices of houses sold in Ames in 2009 exhibits some right skew

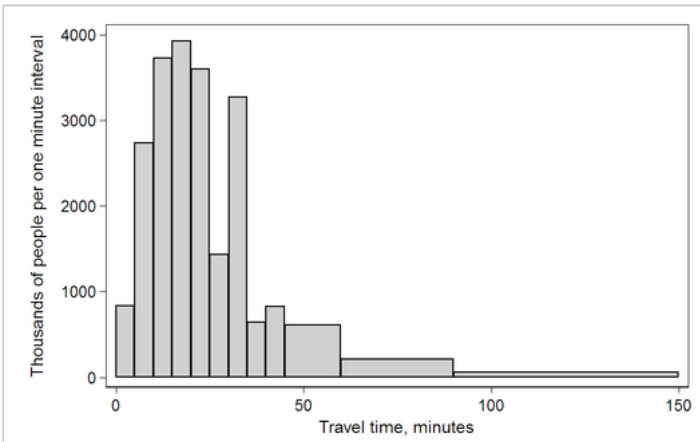


Aces by players in a grand slam tennis tournament, facetted by gender. There are more aces in the men's game.

The U.S. Census Bureau found that there were 124 million people who work outside of their homes.[8] Using their data on the time occupied by travel to work, the table below shows the absolute number of people who responded with travel times "at least 30 but less than 35 minutes" is higher than the numbers for the categories above and below it. This is likely due to people rounding their reported journey time. The problem of reporting values as somewhat arbitrarily rounded numbers is a common phenomenon when collecting data from people.

Data by absolute numbers

Interval	Width	Quantity	Quantity/width
0	5	4180	836
5	5	13687	2737
10	5	18618	3723
15	5	19634	3926
20	5	17981	3596
25	5	7190	1438
30	5	16369	3273
35	5	3212	642
40	5	4122	824
45	15	9200	613
60	30	6461	215
90	60	3435	57



Histogram of travel time (to work), US 2000 census. Area under the curve equals the total number of cases. This diagram uses Q/width from the table.

This histogram shows the number of cases per unit interval as the height of each block, so that the area of each block is equal to the number of people in the survey who fall into its category. The area under the curve represents the total number of cases (124 million). This type of histogram shows absolute numbers, with Q in thousands.

Data by proportion

Interval	Width	Quantity (Q)	Q/total/width
0	5	4180	0.0067
5	5	13687	0.0221
10	5	18618	0.0300
15	5	19634	0.0316
20	5	17981	0.0290
25	5	7190	0.0116
30	5	16369	0.0264
35	5	3212	0.0052
40	5	4122	0.0066
45	15	9200	0.0049
60	30	6461	0.0017
90	60	3435	0.0005

This histogram differs from the first only in the vertical scale. The area of each block is the fraction of the total that each category represents, and the total area of all the bars is equal to 1 (the fraction meaning "all"). The curve displayed is a simple density estimate. This version shows proportions, and is also known as a unit area histogram.

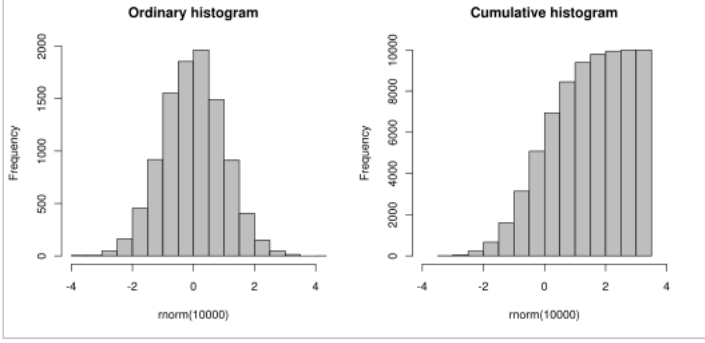
In other words, a histogram represents a frequency distribution by means of rectangles whose widths represent class intervals and whose areas are proportional to the corresponding frequencies: the height of each is the average frequency density for the interval. The intervals are placed together in order to show that the data represented by

the histogram, while exclusive, is also contiguous. (E.g., in a histogram it is possible to have two connecting intervals of 10.5–20.5 and 20.5–33.5, but not two connecting intervals of 10.5–20.5 and 22.5–32.5. Empty intervals are represented as empty and not skipped.)[9]

Mathematical definition

In a more general mathematical sense, a histogram is a function m_i that counts the number of observations that fall into each of the disjoint categories (known as *bins*), whereas the graph of a histogram is merely one way to represent a histogram. Thus, if we let n be the total number of observations and k be the total number of bins, the histogram m_i meets the following conditions:

$$n = \sum_{i=1}^{k} m_i.$$



An ordinary and a cumulative histogram of the same data. The data shown is a random sample of 10,000 points from a normal distribution with a mean of 0 and a standard deviation of 1.

Cumulative histogram

A cumulative histogram is a mapping that counts the cumulative number of observations in all of the bins up to the specified bin. That is, the cumulative histogram M_i of a histogram m_j is defined as:

$$M_i = \sum_{j=1}^{i} m_j.$$

Number of bins and width

There is no "best" number of bins, and different bin sizes can reveal different features of the data. Grouping data is at least as old as Graunt's work in the 17th century, but no systematic guidelines were given[10] until Sturges's work in 1926.[11]

Using wider bins where the density is low reduces noise due to sampling randomness; using narrower bins where the density is high (so the signal drowns the noise) gives greater precision to the density estimation. Thus varying the bin-width within a histogram can be beneficial. Nonetheless, equal-width bins are widely used.

Some theoreticians have attempted to determine an optimal number of bins, but these methods generally make strong assumptions about the shape of the distribution. Depending on the actual data distribution and the goals of the analysis, different bin widths may be appropriate, so experimentation is usually needed to determine an appropriate width. There are, however, various useful guidelines and rules of thumb.[12]

The number of bins k can be assigned directly or can be calculated from a suggested bin width h as:

$$k = \left\lceil \frac{\max x - \min x}{h} \right\rceil.$$

The braces indicate the ceiling function.

Square-root choice

$$k = \sqrt{n},$$

which takes the square root of the number of data points in the sample (used by Excel histograms and many others).[13]

Sturges' formula

Sturges' formula[11] is derived from a binomial distribution and implicitly assumes an approximately normal distribution.

$$k = \lceil \log_2 n \rceil + 1,$$

It implicitly bases the bin sizes on the range of the data and can perform poorly if $n < 30$, because the number of bins will be small—less than seven—and unlikely to show trends in the data well. It may also perform poorly if the data are not normally distributed.

Rice Rule

$$k = \lceil 2n^{1/3} \rceil,$$

The Rice Rule [14] is presented as a simple alternative to Sturges's rule.

Doane's formula

Doane's formula[15] is a modification of Sturges' formula which attempts to improve its performance with non-normal data.

$$k = 1 + \log_2(n) + \log_2\left(1 + \frac{|g_1|}{\sigma_{g_1}}\right)$$

where g_1 is the estimated 3rd-moment-skewness of the distribution and

$$\sigma_{g_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}$$

Scott's normal reference rule

$$h = \frac{3.5\hat{\sigma}}{n^{1/3}},$$

where $\hat{\sigma}$ is the sample standard deviation. Scott's normal reference rule[16] is optimal for random samples of normally distributed data, in the sense that it minimizes the integrated mean squared error of the density estimate.[10]

Freedman–Diaconis' choice

The Freedman–Diaconis rule is:[17][10]

$$h = 2\frac{\mathrm{IQR}(x)}{n^{1/3}},$$

which is based on the interquartile range, denoted by IQR. It replaces 3.5σ of Scott's rule with 2 IQR, which is less sensitive than the standard deviation to outliers in data.

Minimizing cross-validation estimated squared error

This approach of minimizing integrated mean squared error from Scott's rule can be generalized beyond Normal distributions, by using leave-one out cross validation:[18][19]

$$\arg\min_h \hat{J}(h) = \arg\min_h \left(\frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h} \sum_k N_k^2 \right)$$

Here, N_k is the number of datapoints in the *k*th bin, and choosing the value of *h* that minimizes *J* will minimize integrated mean squared error.

Choice based on minimization of an estimated L^2[20] risk function

$$\arg\min_h \frac{2\bar{m} - v}{h^2}$$

where $\bar{m}$ and v are mean and biased variance of a histogram with bin-width h, $\bar{m} = \frac{1}{k}\sum_{i=1}^{k} m_i$ and $v = \frac{1}{k}\sum_{i=1}^{k}(m_i - \bar{m})^2$.

Remark

A good reason why the number of bins should be proportional to $n^{1/3}$ is the following: suppose that the data are obtained as n independent realizations of a bounded probability distribution with smooth density. Then the histogram remains equally »rugged« as n tends to infinity. If s is the »width« of the distribution (e. g., the standard deviation or the inter-quartile range), then the number of units in a bin (the frequency) is of order nh/s and the *relative* standard error is of order $\sqrt{s/(nh)}$. Comparing to the next bin, the relative change of the frequency is of order h/s provided that the derivative of the density is non-zero. These two are of the same order if h is of order $s/n^{1/3}$, so that k is of order $n^{1/3}$. This simple cubic root choice can also be applied to bins with non-constant width.

See also

- Data binning
- Density estimation
  - Kernel density estimation, a smoother but more complex method of density estimation
- Entropy estimation
- Freedman–Diaconis rule
- Image histogram

Wikimedia Commons has media related to ***Histograms***.

- Pareto chart
- Seven Basic Tools of Quality
- V-optimal histograms

References

1. Pearson, K. (1895). "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. **186**: 343–414. Bibcode:1895RSPTA.186..343P (http://adsabs.harvard.edu/abs/1895RSPTA.186..343P). doi:10.1098/rsta.1895.0010 (https://doi.org/10.1098%2Frsta.1895.0010).
2. Howitt, D. and Cramer, D. (2008) *Statistics in Psychology*. Prentice Hall
3. Freedman, D. Pisani, R. and Purves, R. 1998. Statistics (Third edition). W.W.Norton
4. Charles Stangor (2011) "Research Methods For The Behavioral Sciences". Wadsworth, Cengage Learning. ISBN 9780840031976.
5. David W. Scott (December 2009). "Averaged shifted histogram" (https://www.researchgate.net/publication/229760716_Averaged_shifted_histogram). *Wiley Interdisciplinary Reviews: Computational Statistics*. **2:2**: 160–164. doi:10.1002/wics.54 (https://doi.org/10.1002%2Fwics.54).
6. Nancy R. Tague (2004). "Seven Basic Quality Tools" (http://www.asq.org/learn-about-quality/seven-basic-quality-tools/overview/overview.html). *The Quality Toolbox*. Milwaukee, Wisconsin: American Society Quality. p. 15. Retrieved 2010-02-05.
7. M. Eileen Magnello (December 2006). "Karl Pearson and the Origins of Modern Statistics: An Elastician becomes a Statistician" (http://www.rutherfordjournal.org/article010107.html). *The New Zealand Journal for the History and Philosophy of Science and Technology*. 1 volume. OCLC 682200824 (https://www.worldcat.org/oclc/682200824).
8. US 2000 census (http://www.census.gov/prod/2004pubs/c2kbr-33.pdf).
9. Dean, S., & Illowsky, B. (2009, February 19). Descriptive Statistics: Histogram. Retrieved from the Connexions Web site: http://cnx.org/content/m16298/1.11/
10. Scott, David W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley.
11. Sturges, H. A. (1926). "The choice of a class interval". *Journal of the American Statistical Association*: 65–66. doi:10.1080/01621459.1926.10502161 (https://doi.org/10.1080%2F01621459.1926.10502161). JSTOR 2965501 (https://www.jstor.org/stable/2965501).
12. *e.g.* § 5.6 "Density Estimation", W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S* (2002), Springer, 4th edition. ISBN 0-387-95457-0.
13. "EXCEL Univariate: Histogram" (http://cameron.econ.ucdavis.edu/excel/ex11histogram.html).
14. Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University (chapter 2 "Graphing Distributions", section "Histograms")
15. Doane DP (1976) Aesthetic frequency classification. American Statistician, 30: 181–183
16. Scott, David W. (1979). "On optimal and data-based histograms". *Biometrika*. **66** (3): 605–610. doi:10.1093/biomet/66.3.605 (https://doi.org/10.1093%2Fbiomet%2F66.3.605).
17. Freedman, David; Diaconis, P. (1981). "On the histogram as a density estimator: L_2 theory". *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*. **57** (4): 453–476. doi:10.1007/BF01025868 (https://doi.org/10.1007%2FBF01025868).
18. Wasserman, Larry (2004). *All of Statistics*. New York: Springer. p. 310. ISBN 978-1-4419-2322-6.
19. "Optimizing the binwidth for the histogram using cross validation - Maikol Solís" (http://maikolsolis.com/optimizing-histogram-cross-validation/).
20. Shimazaki, H.; Shinomoto, S. (2007). "A method for selecting the bin size of a time histogram" (http://www.mitpressjournals.org/doi/abs/10.1162/neco.2007.19.6.1503). *Neural Computation*. **19** (6): 1503–1527. doi:10.1162/neco.2007.19.6.1503 (https://doi.org/10.1162%2Fneco.2007.19.6.1503). PMID 17444758 (https://www.ncbi.nlm.nih.gov/pubmed/17444758).

Further reading

- Lancaster, H.O. *An Introduction to Medical Statistics.* John Wiley and Sons. 1974. ISBN 0-471-51250-8

External links

- Journey To Work and Place Of Work (http://www.census.gov/population/www/socdemo/journey.html) *(location of census document cited in example)*
- Smooth histogram for signals and images from a few samples (http://www.mathworks.com/matlabcentral/fileexchange/30480-histconnect)
- Histograms: Construction, Analysis and Understanding with external links and an application to particle Physics. (http://quarknet.fnal.gov/toolkits/ati/histograms.html)
- A Method for Selecting the Bin Size of a Histogram (http://2000.jukuin.keio.ac.jp/shimazaki/res/histogram.html)
- Histograms: Theory and Practice (http://www.stat.rice.edu/~scottdw/stat550/HW/hw3/c03.pdf), some great illustrations of some of the Bin Width concepts derived above.
- Histograms the Right Way (http://www.astroml.org/user_guide/density_estimation.html)
- Interactive histogram generator (http://www.shodor.org/interactivate/activities/histogram/)
- Matlab function to plot nice histograms (http://www.mathworks.com/matlabcentral/fileexchange/27388-plot-and-compare-nice-histograms-by-default)
- Dynamic Histogram in MS Excel (http://excelandfinance.com/histogram-in-excel/)
- Histogram construction (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ModelerActivities_MixtureModel_1) and manipulation (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_PowerTransformFamily_Graphs) using Java applets, and charts (http://www.socr.ucla.edu/htmls/SOCR_Charts.html) on SOCR

Categories: Statistical charts and diagrams | Quality control tools | Estimation of densities | Nonparametric statistics | Frequency distribution

---