REVISITING HISTOGRAM

Demudu Naganaidu (GS49320)
Supervisory committee:
Assoc Prof. Dr. Mohd Bakri Adam
Assoc Prof. Dr. Jayanthi Arrasan
Dr Iskandar Bin Ishak

Institute of Mathematical Research
University Putra Malaysia

15th May 2018

Outline of Presentation

Introduction

- Exploratory Data Analysis (EDA) is an approach of analyzing data visually without a prior assumptions on parametric model of the data, error terms, outliers, modality and relationship with other variables.
- EDA helps the researchers to model data based on the what is revealed through exploring with various graphical methods.
- Histogram is one of the important tools used in EDA to summaries large amount data and to visualise the distribution of data.
- It is a nonparametric density estimator.

Introduction(cont..)

- Once a histogram is constructed, general attributes of the data such as symmetry, modality, central location and spread of the data can be revealed.
- No matter how a histogram created the bins size must be decided. Bins can be either all same size (i.e same width) or different size.
- Too many bin makes the histogram uneven and unable to find the underlying trend. Too few bin gives little information about the data.

Introduction(cont..)

- There is no "the best" number of bins, as different bin sizes can reveal different features of the data.
- One usually try different bin numbers, before choosing one that illustrate the salient features of the data.
- The number of bins k for hisogram with equal width can be determined from a suggested bin width h or vice versa.

$$k = \left\lceil \frac{Max(data) - Min(data)}{h} \right\rceil$$

Literature Review I

- (Sturges, 1926) is one of the most cited scholar when it comes to histogram bin selection. He claims proper distribution is distributed into bins by series of binomial coefficients. He presented the number of bins, k for a dataset with N observations was written as:

$$k = 1 + 3.322\ log(N) \tag{1}$$

The bin width, h can be computed as:

$$h = \frac{R}{k} \tag{2}$$

where R is range of data.

Literature Review II

- Sturges' suggest that suitable bin width should be 1,2,5,10,20 or etc so that the theoretical bin width, h computed from (2) can approximated to next smaller convenient bin width.
- Sturges' rule only suitable for strictly normal data, argued (Scott, 2009) . He points out that data from other type of distribution require more bins.
- Scott previously in 1979 introduced a new formula for finding the bin width (Scott, 1979) by including the standard deviation to address problems of bias and variance in estimation.

Literature Review III

- He propose new method to construct histogram using mean squared error criteria that is more rationale according to him. Mean squared error (MSE) of histogram estimate, $\hat{f}(x)$, of the true density value, $f(x)$, defined by

$$MSE(x) = E\left\{\hat{f}(x) - f(x)\right\}^2 \tag{3}$$

Scott derived a general term for the width of the histogram as follows:

$$h^*{}_n = \left\{\frac{6}{\int_{-\infty}^{\infty} f'(x)^2 dx}\right\}^{1/3} n^{-1/3} \tag{4}$$

Literature Review IV

and for normal data, $h^*_n = 2 \times 3^{\frac{1}{3}} \pi^{\frac{1}{6}} \sigma n^{\frac{-1}{3}}$
when σ is unknown the estimate from sample, sample standard deviation is used giving the bin width,
$h = 3.49sn^{-1/3}$

- (Freedman & Diaconis, 1981) came out with a robust method using the Inter Quartile Range, IQR, where the bin width $h^* = 2(IQR)n^{\frac{-1}{3}}$
- (Scott, 1985) proposed the average shifted histogram as variation in density estimation. An algorithm by choosing m histograms but with different bin locations and averaging them to get average shifted histogram.

Literature Review V

- (Wand, 1997) regards that the bin width is the most important parameter when it comes to histograms construction. Whether a histogram is 'over smooth' or 'under smooth' is controlled by this parameter. He extended the Scott's rule to provide a simple and with asymptotic performance. However the method proposed is not straight forward.

Literature Review VI

- (Bura, Zhmurov, & Barsegov, 2009) introduced the cross validation (CV) method and derived:

$$CV\,[h] = \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)} \sum_{k=1}^{m} \frac{n_k^2}{n^2} \tag{5}$$

where m is number of bins, $h = (x_{max} - x_{min})/m$, and n_k is the kth bin count. The value of h_i that corresponds to the minimum of CV[h] defines the h_{opt}, optimum bin width.

- Variation from Sturges formula introduced by Doane in 1976 (Wand, 1997)

$$h = 1 + \frac{log(n)+log(1+c_1)}{log(2)} \text{ where}$$

Literature Review VII

$$c_1 = \frac{m_3}{m_2^{3/2}} \left[\frac{(n+1)(n+3)}{6(n-2)} \right]^{1/2} \text{ and}$$

$$m_j = \sum_{i=1}^{n} (X_i - \hat{X})^{j/n}$$

Problem Statement

- In constructing histogram, important parameters are: (i) number of bins or intervals, (ii) bin width and (iii) lower limit of first bin (Waterman & Whiteman, 1978).
- In statistical theory however only few guidelines are available for selecting the number of bins or bin width for the histogram (He & Meeden, 1997; Birgé & Rozenholc, 2006)
- Despite some suggested methods to determine the number of bins, there is no single method is agreed upon.
- Statistical softwares available uses different methods or modified the existing methods so that the bin width have nice break points.

Problem Statement

- S-Plus uses Sturges Rule but with modification (Wand, 1997). Similarly R Programming.
- STATA gives a range for the user to select the bin width between Sturges, Scott, Freedman-Diaconis and others.
- Inexperienced researches may find it challenging to decide which is the best method.
- The central tendency measures and variance calculated from raw data can used in deciding the number bins.
- Detecting the outliers from histogram not very helpful due the method of constructing. The modification of construction histogram can reveal the outliers more effectively.

Research Aims and Objectives

This study aims to achieve following objectives:

- Propose a method using raw data central tendency measures and variation to decide the number of bins.
- Evaluate performance of new method with existing methods
- Propose a new method for construction of histogram to detect outliers
- Evaluate performance of new method against method using Inter Quartile Range (IQR) to detect outliers
- Propose a fixed frequency histogram for the purpose of segmentation or classification.

Methodology
Frequency Table

- Consider a continuous data set $x_1, x_2, x_3, ..., x_n$ with unknown density f where all values are in an interval [a,b). Group the data into k bins with end points

$$a = a_0 < a_1, < a_2 < ... < a_k = b$$

- Let the frequency of data in each bin B_j be denoted by f_j and with fixed bin width $h = a_j - a_{j-1}$.
- Each bin range can be written as :

$$B_j = [a_0 + (j-1)h, a_0 + jh), j = 1, 2, ..., k$$

$$f_j = \sum_{i=1}^{n} I(x_i \in B_j), \text{ and } \sum f_j = n$$

- Denote m_j as the center of the bin B_j

Methodology
Frequency Table

- The data with n observations then can be summarised in frequency table.

Table: Frequency Table

	Mark	Frequency	Relative Frequency
B_j	m_j	f_j	$\frac{f_j}{n}$
B_1	m_1	f_1	$\frac{f_1}{n}$
B_2	m_2	f_2	$\frac{f_2}{n}$
.	.	.	.
.	.	.	.
B_k	m_k	f_k	$\frac{f_k}{n}$
		$\sum = n$	$\sum = 1$

Methodology
Histogram

- A graphical display of frequency table is called as Histogram. There are two types of histogram: (i) Frequency Histogram and (ii) Relative Frequency Histogram
- To visualise Frequency Hsitogram plot the frequency as the height of a column, with the width of the column representing the width of bin.
- Relative Frequency Histogram are obtained by representing the height of the bin by relative frequency.
- The histogram in density form is defined as

$$\hat{f}(x) = \frac{f_j}{nh} \text{ where } \hat{f}(x) \geq 0 \text{ and } \int \hat{f}(x)dx = 1$$

Methodology
Measures of Central Tendency

- Mean

$$\bar{x} = \frac{\sum f_j m_j}{n}$$

- Median

$$Median = L_1 + \left[\frac{\frac{n}{2} - cf}{f} * h\right], \text{ where}$$

L_1 = lower limit of median bin
cf =the cumulative frequency of the bin preceding the median bin
h = bin width
median bin is the bin where the item $\frac{n}{2}$ located.

Methodology
Measures of Central Tendency

- Mode

$$Mode = L_1 + \left[\frac{fm-fb}{(fm-fb)+(fm-fa)} * h\right], \text{ where}$$

L_1 = lower limit of modal bin
fm = frequency of modal bin
fb = frequency of bin before modal bin
fa = frequency of bin after modal bin
h = bin width
modal bin is the bin with highest frequency

Methodology
Variance

- Variance

$$S^2 = \frac{\sum_{j=1}^{k} f_j\left(m_j - \hat{x}\right)^2}{n-1} \text{ , where}$$

$m_j =$ as the center of bin B_j
$f_j =$ frequency of bin B_j
$j = 1, 2, ..k$

Methodology
New Method for Histogram Binning

- Bins will be useful to calculate averages, variance, skewness and other moments of frequency distributions formed with the k number of bins (Sturges, 1926) .
- Number of bins between 5 to 20 is adequate for real set of data (Scott, 1979)
- As we are using the histogram as non parametric density estimator, the central tendency measures and variance obtained from histogram must be close to the one obtained from raw data.

Methodology
New Method for Histogram Binning

- To do this an iteration method is proposed starting with 5 bins. Based on the 5 bins, data is grouped and respective frequencies to be obtained. Central tendency and variance from grouped data then compared with the same measures from raw data.
- Next is to increase the bins by 1 and repeat same process above. Once we have measure from 5 bins to 20 bins, comparison can be made to determine the which number of bins provide the closest estimate to the raw data estimate.

Methodology
Detecting Outliers from Histogram

- The current method detecting outliers are from histogram are not effective.
- This is due to the current method of which requires to determine the lower limit of first bin or the upper limit of last bin.
- New method proposed to use the median at starting point to draw histogram on both left and right. The columns of histograms are drawn until the minimum value and maximum value are included in the histogram.
- Data in columns after gaps by two columns can be considered as outliers.

References I

Birgé, L., & Rozenholc, Y. (2006). How many bins should be put in a regular histogram. *ESAIM: Probability and Statistics*, *10*, 24–45.

Bura, E., Zhmurov, A., & Barsegov, V. (2009). Nonparametric density estimation and optimal bandwidth selection for protein unfolding and unbinding data. *The Journal of chemical physics*, *130*(1), 01B602.

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L 2 theory. *Journal of Probability Theory and Related Areas*, *57*(4), 453-476.

References II

He, K., & Meeden, G. (1997). Selecting the number of bins in a histogram: A decision theoretic approach. *Journal of Statistical Planning and inference*, *61*(1), 49–59.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66*(3), 605–610.

Scott, D. W. (1985, 09). Averaged shifted histograms: Effective nonparametric density estimators in several dimensions. *Ann. Statist.*, *13*(3), 1024–1040. Retrieved from `https://doi.org/10.1214/aos/1176349654` doi: 10.1214/aos/1176349654

Scott, D. W. (2009). Sturges' rule. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(3), 303–306.

References III

Sturges, H. A. (1926). The choice of a class interval. *Journal of the american statistical association*, *21*(153), 65–66.

Wand, M. (1997). Data-based choice of histogram bin width. *The American Statistician*, *51*(1), 59–64.

Waterman, M., & Whiteman, D. (1978). Estimation of probability densities by empirical density functions. *International Journal of Mathematical Education in Science and Technology*, *9*(2), 127–137.