Some Remarks on Class Interval of Histograms

Yasuhisa HIRAI

1. Introduction

Since the well-known formula by Sturges, several methods have been proposed for finding the optimal class width in drawing a histogram, and each of which is based on different criterion. In this paper we propose a method for choice of class interval in the case that data has normal distribution. Optimal class interval from our method is compared with Mori's method.

2. Short review on several methods

Stürges(1932) gave a formula

Number of class \( k = 1 + \log n \), class interval \( w = R/(1 + \log n) \), where \( n \) is the number of data and \( R \) is the range of data, from the binomial distribution of data.

Similar formula,

\[
k = 1 + 2.2 \log_{10} n
\]

was given by Larson(1975) as a first choice.

Doane(1976) modified Sturges' formula considering the case that the distribution is not symmetric.

\[
2^{a-1} = n \left( 1 + \frac{g}{6(n-2)} \right)
\]

where \( g \) is the skewness of data.

Suzuki(1985) used a rounded number for the class interval and also the endpoint of a class.

\[
b = \left[ \log_{10} R/4 \right]
\]

\[
a(R) = 10^{\alpha} \cdot \left( 1 + \sqrt{(R/4 \cdot 10^a - 1)^2/10} \right)
\]

\[
w = a \times (\text{integer})
\]

\[
l = a \times (\text{integer}) \text{ or } l = (\text{integer}) + a/2, l \leq \text{minimum} \text{ data} \leq l + w
\]

Mori(1974) proposed the following method by minimizing the mean squared error of a histogram estimate \( f_n \) of the true density \( f(x) \).

Letting \( N_j \) denote the frequency of data in the interval \( I_j = [j w, (j + 1) w] \) and \( f_n(x) = \frac{N_j}{nw} \cdot x \in I_j \), the mean squared error

\[
\varphi n(w) = E \cdot \int_{-\infty}^{\infty} (f_n(x) - f(x))^2 dx
\]

is used here as an index of approximation of \( f_n \) to \( f \). A class interval \( w \) which minimizes the \( \varphi_n \) is the optimal \( w \). He finally derived the formula for the optimal class interval

\[
w = \left( \frac{6}{Bn} \right)^{1/n}, \quad B = \int_{-\infty}^{\infty} f(x) dx.
\]
This gives \( w \sim 3.49 \sigma n^{-1/2} \) in the case that \( f(x) \) has the normal distribution. Exactly the same result was given by Scott (1979) with the same purpose but a different proof.

Among other formulas,

\[ K = \sqrt{n} \]

is a simple one.

3. An index for the optimal histogram

We here assume that the given data has the normal distribution. An index for an approximation of a histogram to the density function is defined as follows.

Let \( k \) denote the number of the classes in a histogram of \( n \) samples, \( x_0 \) be the (left-hand) endpoint, and \( x_1, x_2, ..., x_k \) be other endpoints. Let \( N_i \) denote the frequency of data \( x, x_{i-1} < x \leq x_i \). Let \( f_\ast = N_i / (nw) \), where \( w = x_i - x_{i-1} \), so that the sum of the area of all columns of the histogram adds to unity. We here take the symmetry of the normal density into consideration, so only two ways of class location are provided, i.e. 1) the sample mean \( \bar{x} \) is the endpoint of a class and 2) \( \bar{x} \) is located in the middle point of a class.

The area difference \( D \) between the histogram and the density function is

\[
D = \int_{-\infty}^{\infty} | f(x) - f_\ast | \, dx
\]

\[
= \int_{-\infty}^{\infty} \left| (\sqrt{2 \pi \sigma})^{-1} \exp\left(-\frac{(x - \mu)^2}{2 \sigma^2}\right) - f_\ast \right| \, dx.
\]

Histograms with smaller \( D \) should have the shapes closer to the density function. Therefore we let this \( D \) be our index for the optimal histogram i.e. optimal choice of \( w \).

3.1 Exploration by simulation

The simulation study gave optimal class intervals from the normal random numbers (\( n = 100, 200, ..., \)). Fig.1a shows the plot of \( w \)'s and \( D \)'s from 300 normal random numbers. In the figure, the cross and the circle is from the first and the second way of class location respectively. The optimal case is obtained in the bottom, (\( w \) is around 0.45). The simulation results for bigger \( n \)'s are shown in Fig.1b(\( n = 500 \)), Fig.1c(\( n = 1000 \)), and Fig.1d(\( n = 2000 \)). With bigger \( n \), we have more clear shape in plotting and smaller \( w \) as an optimal one. But these figures show that optimal \( w \)'s seem to be found in some range.
Some Remarks on Class Interval of Histograms

Figure 1

3.2 Numerical comparison with Mori's methods

Histograms are actually drawn for 2000 normal random numbers with optimal \( w \) by Mori's method(Fig.2a) and by our proposed method(Fig.2b). In this case Mori's histogram looks less smooth, probably due to too many classes.
3.3 An optimal number of class

In the case of actual data, to know the number of the class $k$ is often more useful than the class interval $w$. Fig. 3 shows the plot of $n$ (= number of data) and $k$ from our method and other several methods are superimposed for comparison. In the figure, $\square$ is from our method, $S$ from Sturges, $R$ from $k = \sqrt{n}$, and $M$ from Mori's method. Up to around $n = 600$, the optimal $k$'s from our method are similar to the ones from Sturges, but especially if $n > 600$ optimal $k$'s from our methods exist in a wide range.
Some Remarks on Class Interval of Histograms

References


(Received July 8, 1989)