What Is the Optimal Bin Size of a Histogram: An Informal Description

Afshin Gholamy[1] and Vladik Kreinovich[2]
[1]Department of Geological Sciences
[2]Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
afshingholamy@gmail.com, vladik@utep.edu

Abstract

A natural way to estimate the probability density function of an unknown distribution from the sample of data points is to use histograms. The accuracy of the estimate depends on the size of the histogram's bins. There exist heuristic rules for selecting the bin size. In this paper, we show that these rules indeed provide the optimal value of the bin size.

1 Formulation of the Problem

Need to estimate pdfs. One of the most frequent ways to describe a probability distribution is by specifying its probability density function (pdf)

$$\rho(x) \stackrel{\text{def}}{=} \frac{dp}{dx} = \lim_{h \to 0} \frac{\text{Prob}(X \in [x, x+h])}{h}.$$

In many practical situations, all we know about a probability distribution is a sample of data points corresponding to this distribution. How can we estimate the pdf based on this sample?

Enter histograms. A natural way to estimate the limit when h tends to 0 is to consider the value of the ratio corresponding to some small h:

$$\rho(x) \approx \frac{\text{Prob}(X \in [x, x+h])}{h}.$$

To use this expression, we need to approximate the corresponding probabilities $\text{Prob}(X \in [x, x+h])$. By definition, the probability of an event is the limit of this event's frequency when the number of data points increases. In particular,

$$\text{Prob}(X \in [x, x+h]) = \lim_{n \to \infty} \frac{n([x, x+h])}{n},$$

where n is the overall number of data points and $n([x, x+h])$ denotes the number of data points within the interval $[x, x+h]$. Thus, as an estimate for the corresponding probability, we can get the frequency $f([x, x+h])$ of this event, i.e., the ratio

$$f([x, x+h]) \approx \frac{n([x, x+h])}{n}.$$

This idea leads to the following estimate for $\rho(x)$;

$$\rho(x) \approx \frac{f([x, x+h])}{h}.$$

This estimate, known as a *histogram* approximation, was first introduced by Karl Pearson in [5]; for details, see, e.g., [2, 3, 4].

In a histogram, the range of possible values of the corresponding quantity x is divided into intervals $[x_i, x_{i+1}]$ $(1 \le i \le k)$ called *bins*. In most practical cases, all the bins have the same width $h_i = x_{i+1} - x_i$, i.e., $h_1 = \ldots = h_k = h$ for some $h > 0$.

For each bin i, we then estimate (and plot) the frequency $f_i \stackrel{\text{def}}{=} \dfrac{n([x_i, x_{i+1}])}{n}$ with which the data points fall into this bin.

For values x from the corresponding interval, the probability density

$$\rho(x) \stackrel{\text{def}}{=} \frac{dp}{dx} = \lim_{h \to 0} \frac{\text{Prob}(X \in [x, x+h])}{h}$$

is approximated as the ratio

$$\rho_i = \frac{f_i}{h_i}.$$

Need to select a bin size. To form a histogram, we need to select the bin size h.

How bin sizes are selected now. In situations when we have an additional information about the corresponding probability distribution, we can formulate the bin selection problem as a precise optimization problem, and get the solution; see, e.g., [1, 6]. In many such cases, the optimal bin size h_{opt} decreases with the number n of data points as

$$h_{\text{opt}} = \text{const} \cdot \frac{s}{n^{1/3}},$$

where s is the "width" of the distribution – this can be the range of the interval at which $\rho(x)$ is positive, or, for distributions like Gaussian for which the pdf is never equal to 0, the difference between two quantiles.

In many practical situations, however, we do not have any additional information about the probability distribution – only the sample itself. In such situations, several heuristic rules have been proposed, most of them using the same dependence $h \sim \dfrac{s}{n^{1/3}}$; see, e.g., [2, 3, 4]. These heuristic rules are justified by three things:

- first, as we have mentioned, under certain conditions, these rules do provide provably optimal bin sizes; so it makes sense to assume that they are optimal in more general situations as well;
- second, the experience of using these rules shows that they, in general, work better than several previously proposed heuristic rules,
- third, under these rules, the two components of the pdf estimation error are approximately equal to each other, and equality of two error components is often an indication of optimality.

What we do in this paper. In this paper, we provide a somewhat stronger justification for the existing heuristic methods of selecting the bin size. Specifically, we provide (informal) arguments that the current heuristic rules indeed provide the optimal bin size – namely, the bin size for which the pdf approximation error is the smallest possible.

2 Which Bin Sizes Are Optimal: Analysis of the Problem and the Resulting Recommendation

Two reasons why a histogram is different from the pdf. To find the optimal bin size h_{opt}, we need to describe how the approximation error – i.e., the error with which the histogram approximates the actual pdf – depends on the bin size h. There are two reasons why the histogram is different from the pdf – and these reasons lead to two components of the approximation error:

- first, for each bin, for all the values x from this bin, the histogram provides the same value while the probability density function $\rho(x)$ has, in general, different values at different points x inside this bin;
- second, each estimate ρ_i is based on a finite sample, and it is well known that in statistics, estimates based on a finite sample are approximate.

Let us therefore estimate both components of the approximation error.

First component of the approximation error: approximation error caused by the finiteness of the bin size. The first error component is caused by the fact that for all points x from the i-th bin, we use the same approximating value ρ_i, while the actual pdf $\rho(x)$ is, in general, different for different points within the bin.

We can describe the corresponding approximation error as follows: instead of using the values $\rho(x)$ corresponding to different points x from the bin, we select one point x_0 from the bin, and use the value $\rho(x_0)$ instead of the actual value $\rho(x)$.

The larger the distance $|x - x_0|$ between the points x and x_0, the more the actual value $\rho(x)$ is different from our approximation $\rho(x_0)$. The worst case is

when the difference $|x - x_0|$ is the largest possible. Thus, as x_0, we should select the point for which this largest distance is as small as possible.

One can easily check that this means selecting the midpoint $x_0 = x_{\rm mid}$ of the bin. Indeed, in this case, the worst-case distance is equal to $s/2$, while if we select the point x_0 tilted to the left or to the right, the worst-case distance will be larger.

So, the first component of the approximation error comes from the difference $|\rho(x) - \rho(x_{\rm mid})|$ between the values of the pdf at the points x and $x_{\rm mid}$ for which $|x - x_{\rm mid}| = h/2$.

How big is this difference? Most practical distributions are unimodal: the corresponding pdf starts from 0, increases until it reaches its maximum value, and then decreases back to 0.

On the interval of the distribution width s, the pdf $\rho(x)$ goes from 0 to its maximum value $\rho_{\max}$ and back. We do not know which part of the interval of size s corresponds to increasing and which to decreasing. Since there is no reason to believe that the increasing part is longer than the decreasing one or vice versa, it make sense to assume, in our estimates, that these two parts has the same width, i.e., that the pdf increases on the interval of width $s/2$ and then decrease on the interval of the same width.

On the interval of size $s/2$, the value of the pdf $\rho(x)$ increases from 0 to its maximum value $\rho_{\max}$. The change of $\rho(x)$ on the interval of width $h/2$ should be proportional to this wirth, i.e., we should have

$$\Delta\rho \stackrel{\rm def}{=} |\rho(x) - \rho(x_{\rm mid})| \approx \frac{h/2}{s/2} \cdot \rho_{\max}.$$

Thus, the relative value $\dfrac{\Delta\rho}{\rho}$ of the first component of the approximation error is approximately equal to

$$\frac{h/2}{s/2} = \frac{h}{s}. \tag{1}$$

Second component of the approximation error: approximation error caused by the finiteness of the sample. The second component of the approximation error is cased by the fact that each estimate ρ_i is based on the finite sample. It is known (see, e.g., [2, 3, 4]) that when we estimate a parameter based on a sample of size m, we get an estimate with a relative error $\dfrac{1}{\sqrt{m}}$.

On the range of width s we have several bins of size h. Thus, the overall number of bins is equal to $k = s/h$. The overall number of data points is n, so in each of the k bins, we have, on average,

$$m = \frac{n}{k} = \frac{n}{s/h} = \frac{n \cdot h}{s}$$

points. Based on these number of points, we get the following formula for the relative value of the second component of the approximation error:

$$\frac{1}{\sqrt{m}} = \frac{\sqrt{s}}{\sqrt{n \cdot h}}. \tag{2}$$

Overall approximation error. By adding the two components (1) and (2) of the approximation error, we get the following expression for the overall relative approximation error E:

$$E = \frac{h}{s} + \frac{\sqrt{s}}{\sqrt{n \cdot h}}. \tag{3}$$

Let us find the optimal bin size. To find the optimal bin size, we differentiate the expression (3) with respect to h and equate the resulting derivative to 0. As a result, we get

$$\frac{1}{s} - \frac{1}{2} \cdot \sqrt{\frac{s}{n}} \cdot h^{-3/2} = 0,$$

hence

$$\frac{1}{2} \cdot \sqrt{\frac{s}{n}} \cdot h^{-3/2} = \frac{1}{s}.$$

If we multiply both sides of this equality by $h^{3/2} \cdot s$, we conclude that

$$h^{3/2} = 2 \cdot \frac{s^{3/2}}{n^{1/2}}.$$

By raising both sides by the power 2/3, we get the formula

$$h_{\text{opt}} = \text{const} \cdot \frac{s}{n^{1/3}}.$$

This is exactly the heuristic rule that we wanted to justify.

3 Acknowledgments

This work was supported in part by the National Science Foundation grant HRD-1242122 (Cyber-ShARE Center of Excellence).

References

[1] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L_2 theory", *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1981, Vol. 57, No. 4, pp. 453–476.

[2] D. Freedman, R. Pisani, and R. Purves, *Statistics*, W. W. Norton, New York, 1998.

[3] D. Howitt and D. Cramer, *Statistics in Psychology*, Prentice Hall, Upper Saddle River, New Jersey, 2008.

[4] H. O. Lancaster, *An Introduction to Medical Statistics*, John Wiley and Sons, New York, 1974.

[5] K. Pearson, "Controbutions to the mathematical theory of evolution.II. Skew variation in homogeneous matreial", *Philosophical Transactions of the Roay Society A: Mathematical, Physical, and Engineering Sciences*, 1895, Vol. 186, pp. 343–414.

[6] D. W. Scott, "Avarages shifted histogram", *Wiley Interdisciplinary Reviews: Computational Statistics*, 2009, Vol. 2, No. 2, pp. 160–164.