On the Maximum Deviation Between the Histogram and the Underlying Density

David Freedman¹⋆ and Persi Diaconis²★★

¹ Statistics Department, University of California, Berkeley, CA, USA
² Statistics Department, Stanford University, Palo Alto, CA, USA

Summary. The limiting joint distribution of the location and size of the maximum deviation between the histogram and the underlying density is derived. For smooth densities, the location and size of the maximum are asymptotically independent. The size has a limiting double-exponential distribution and the location has a limiting normal distribution.

1. Introduction

A sample of size $k$ is drawn from a density $f$.

(1.1) \[ f \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} f(x) \, dx = 1. \]

A histogram of cell width $h$ is used to estimate $f$. How far off is the estimate? Where does the maximum discrepancy occur? The object of this paper is to describe the asymptotic joint distribution of these two variables.

To state a precise result, assume that

(1.2) $f$ has a unique maximum at $x_0$.

Assume too that $f$ is locally quadratic at $x_0$:

(1.3) \[ f(x_0 + x) = f(x_0) + \frac{1}{2} \alpha x^2 + o(x^2) \quad \text{as} \quad x \to 0, \]

where $\alpha$ is negative; write $f''''(x_0) = \alpha$. This does not assume any differentiability; however, if $f$ is smooth, then $\alpha$ is the ordinary second derivative at $x_0$. Finally, assume

(1.4) \[ \sup_{x} \{ f(x_0 + x) : |x| \geq \delta \} < f(x_0) \quad \text{for any} \quad \delta > 0. \]

⋆ Research partially supported by NSF grant MCS-80-02535
★★ Research partially supported by NSF grant MCS-77-16974
For ordinary functions, (1.2) is equivalent to (1.3) and (1.4).

To define the histogram, choose a point $\lambda_0$ with $\lambda_0 \leq x_0 < \lambda_0 + h$. By definition, cell $j$ of the histogram will run from $\lambda_j = \lambda_0 + hj$ to $\lambda_{j+1} = \lambda_0 + h(j+1)$:

$$\text{cell } j = [\lambda_j, \lambda_{j+1}) = [\lambda_0 + hj, \lambda_0 + h(j+1)),$$  

where $j = 0, \pm 1, \pm 2, \ldots$

The data consists of $k$ independent random variables $X_1, X_2, \ldots, X_k$ with common probability density $f$. By definition, $N_j$ is the number of data points falling in cell $j$. Formally,

(1.5) $N_j$ is the number of indices $i = 1, \ldots, k$ with $\lambda_j \leq X_i < \lambda_{j+1}$.

By definition, the histogram is

(1.6) $H(x) = N_j/(kh)$ for $x \in [\lambda_j, \lambda_{j+1})$.

This definition forces the area under $H(x)$ to be 1. Let $p_j$ be the probability of the $j^{th}$ cell:

(1.7) $p_j = \int_{\lambda_j}^{\lambda_{j+1}} f(x) \, dx$.

Define $f_h(x)$ to be $p_j/h$ for $x$ between $\lambda_j$ and $\lambda_{j+1}$.

The difference between the histogram and the density can be decomposed as:

(1.8) $H(x) - f(x) = H(x) - f_h(x) + f_h(x) - f(x)$.

The term $H(x) - f_h(x)$ represents sampling error; $f_h(x) - f(x)$ represents bias. When $h$ is small, sampling error dominates and the distribution of $\sup_x [H(x) - f(x)]$ is the same as the distribution of $\frac{1}{kh} \sup_j (N_j - kp_j)$. For this reason, it is useful to derive the distribution of the location and size of $\sup_j (N_j - kp_j)$. The following growth condition will be needed:

(1.9) $k \to \infty$ and $h \to 0$ in such a way that $k \sqrt{\frac{1}{h} \left( \log \frac{1}{h} \right)^3} \to \infty$.

In the absence of this condition, large-deviations corrections to the central limit theorem become relevant: see [2] for a related discussion. A final burst of notation:

(1.10) $W_h(x) = \left[ 2 \log \frac{1}{h} - 2 \log \log \frac{1}{h} + x \right]^{1/2}$

(1.11) $\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-u^2/2} \, du$,

(1.12) $2 \rho^2 = |f''(x_0)|/f(x_0)$.

The first result can now be stated: