EFFICIENT ALGORITHMS FOR ESTIMATING THE WIDTH OF NEARLY NORMAL DISTRIBUTIONS

Carl W. AKERLOF

Randall Laboratory of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA

Received 16 September 1982

Typical physics data samples often conform to Gaussian distributions with admixtures of more slowly varying backgrounds. Under such circumstances the standard deviation is known to be a poor statistical measure of distribution width. As an alternative, the performance of Gini's mean difference is compared with the standard deviation and the mean deviation. Variants which sum over subsets of all possible pairs are shown to have statistical efficiencies comparable to the mean difference and mean deviation but do not require extensive data storage or a priori knowledge of the sample mean. These statistics are reasonable candidates for monitoring the distribution width of a real time data stream.

In a variety of circumstances, the optimal performance of a system depends on minimizing the width of a distribution derived from experimental data. Operationally these problems are similar to focusing a camera. As an obvious example, charged particle transport systems are normally tuned by monitoring the width of the beam image at a focal point and varying the current in the magnetic optics until the spot size is smallest. At a slightly greater level of abstraction the optimum momentum resolution of a spectrometer is obtained after a number of calibration constants are varied to minimize the apparent momentum spread of a narrow bandwidth test beam.

Common to both of these examples is the existence of a small but finite admixture of background to an otherwise narrow Gaussian distribution. In the case of beam optics there is always a certain level of beam halo from slit scattering and other sources which may be quite erratic in time. Reliable tuning of such a system requires statistical methods that minimize these effects. Unfortunately the most widely used measure of distribution width is the standard deviation which is known to be a very poor statistic [1] in the usual dirty circumstances with which we must contend.

For a sample drawn from a normally distributed population the standard deviation is the optimal measure of distribution width [2] in the sense that the relative statistical uncertainty is least for this particular parameter for a fixed sample size. However as the population distribution departs from normality the large weights associated with observations far from the mean strongly affect the values obtained. The sensitivity of the standard deviation is such that if 0.2% of the sample is drawn from a population with three times greater width the mean deviation defined by

$$d=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\frac{1}{n}\sum_j x_j\right| \tag{1}$$

is a more stable statistic. Needless to say such a small contamination is practically invisible in a sample of reasonable size.

There is quite a large number of alternative techniques which can be employed for better estimates of distribution width. For computation most of these require memory storage for every member of the sample. In the case of the mean deviation, because the sample mean is not known in advance, the absolute deviations cannot be obtained until after the entire sample has been collected. Similar arguments apply for order statistics (such as the semi interquartile range) since the ordering can not be done until all of the sample data is available. For monitoring real time data streams this storage requirement is inconvenient if

0167-5087/83/0000-0000/$03.00

not impossible to satisfy. This paper will describe a statistical measure of width which can be computed with a simplicity comparable to the standard deviation without the extreme sensitivity to non-Gaussian tails.

The origin of these measures is named *Gini's mean deviation* [3] after the Italian statistician although it was apparently known to Helmert [4] and others [5] in the 1870s. It is defined by

$$g=\frac{2}{n(n-1)}\sum_{\substack{1<i\leqslant n\\ 1\leqslant j<i}}|x_i-x_j|, \tag{2}$$

where the sum runs over all possible sample pairs. The virtue of the mean difference is its relative insensitivity to large deviates. This property is shared by the mean deviation but, unlike the mean deviation, the mean difference does not require prior calculation of the sample mean. An obvious disadvantage is the necessity of summing over all $\frac{1}{2}n(n-1)$ sample pairs which greatly inhibits its utility for general use.

Fortunately similar statistics formed from subsets of all possible pairs retain the useful properties given above without drastic loss of statistical efficiency. These variants of the mean difference can be computed for indefinitely long data streams yet require negligible data storage:

$$\begin{aligned}
g_0&=\frac{2}{n}\sum_{i=1}^{n/2}|x_{2i}-x_{2i-1}|, \qquad n \text{ even}\\
g_1&=\frac{1}{n-1}\sum_{i=2}^{n}|x_i-x_{i-1}|\\
g_2&=\frac{1}{2n-3}\left[\sum_{i=2}^{n}|x_i-x_{i-1}|+\sum_{i=3}^{n}|x_i-x_{i-2}|\right]\\
g_3&=\frac{1}{3n-6}\left[\sum_{i=2}^{n}|x_i-x_{i-1}|+\sum_{i=3}^{n}|x_i-x_{i-2}|+\sum_{i=4}^{n}|x_i-x_{i-3}|\right].
\end{aligned} \tag{3}$$

g_0 requires a summation over $n/2$ disjoint sequential pairs. g_1, g_2, and g_3 are averages over all possible pairs within sequential groups of two, three and four sample elements and contain $n-1$, $2n-3$, and $3n-6$ terms respectively. As long as the observations $[x_i]$ are statistically independent the expectation values of these new statistics are all equal:

$$\langle g_0\rangle=\langle g_1\rangle=\langle g_2\rangle=\langle g_3\rangle=\langle g\rangle \tag{4}$$

The computation of the variance of g_i follows directly from the method of Lomnicki [6,7]:

$$\begin{aligned}
\langle g_0^2\rangle&=\frac{1}{n}\left[2\langle(x_i-x_j)^2\rangle+(n-2)\langle|x_i-x_j||x_k-x_l|\rangle\right],\\
\langle g_1^2\rangle&=\frac{1}{(n-1)^2}\left[(n-1)\langle(x_i-x_j)^2\rangle+2(n-2)\langle|x_i-x_j||x_i-x_k|\rangle\right.\\
&\quad\left.+(n-2)(n-3)\langle|x_i-x_j||x_k-x_l|\rangle\right],\\
\langle g_2^2\rangle&=\frac{1}{(2n-3)^2}\left[(2n-3)\langle(x_i-x_j)^2\rangle+4(3n-8)\langle|x_i-x_j||x_i-x_k|\rangle\right.\\
&\quad\left.+2(2n^2-13n+22)\langle|x_i-x_j||x_k-x_l|\rangle\right],\\
\langle g_3^2\rangle&=\frac{1}{(3n-6)^2}\left[(3n-6)\langle(x_i-x_j)^2\rangle+2(15n-52)\langle|x_i-x_j||x_i-x_k|\rangle\right.\\
&\quad\left.+(9n^2-69n+146)\langle|x_i-x_j||x_k-x_l|\rangle\right],
\end{aligned} \tag{5}$$

$$\langle g^2 \rangle = \frac{1}{n(n-1)}\left[2\langle (x_i - x_j)^2\rangle + 4(n-2)\langle |x_i - x_j||x_i - x_k|\rangle + (n-2)(n-3)\langle |x_i - x_j||x_k - x_l|\rangle\right].$$

In each of the above expressions the first and third expectation values are easily obtained from

$$\langle (x_i - x_j)^2 \rangle = 2\sigma^2, \qquad \langle |x_i - x_j||x_k - x_l| \rangle = \langle g \rangle^2. \tag{6}$$

The correlation term can be reduced to the following expression:

$$\langle |x_i - x_j||x_i - x_k| \rangle = \langle D^2(x) \rangle, \tag{7}$$

where

$$D(x) \equiv 2(xF(x) - G(x)) + \langle x \rangle - x, \tag{8}$$

and $F(x)$ and $G(x)$ are related to the differential distribution function, $f(x)$, by:

$$F(x) = \int_a^x f(x)\,\mathrm{d}x, \qquad G(x) = \int_a^x xf(x)\,\mathrm{d}x. \tag{9}$$

The expectation value of g is:

$$\langle g \rangle = \langle D(x) \rangle = 2\langle 2xF(x) - x \rangle. \tag{10}$$

The variances of g_i are easily obtained from the values of $\langle g_i^2 \rangle$ given by eq. (5). In the limit of large sample size

$$\begin{aligned}
&\lim_{n\to\infty} n\,\mathrm{var}(g_0) = 4\sigma^2 - 2\langle g \rangle^2,\\
&\lim_{n\to\infty} n\,\mathrm{var}(g_1) = 2\sigma^2 + 2\langle D^2 \rangle - 3\langle g \rangle^2,\\
&\lim_{n\to\infty} n\,\mathrm{var}(g_2) = \tfrac{1}{2}(2\sigma^2 + 6\langle D^2 \rangle - 7\langle g \rangle^2),\\
&\lim_{n\to\infty} n\,\mathrm{var}(g_3) = \tfrac{1}{3}(2\sigma^2 + 10\langle D^2 \rangle - 11\langle g \rangle^2),\\
&\lim_{n\to\infty} n\,\mathrm{var}(g) = 4\langle D^2 \rangle - 4\langle g \rangle^2.
\end{aligned} \tag{11}$$

The statistical performance of the mean difference can now be quantitatively compared with other measures such as the standard deviation and mean deviation. For these purposes we will assume for the differential distribution function a sum of two normal distributions with identical mean but different width:

$$f(x) = (1-\epsilon)\phi(x, \sigma_1) + \epsilon\phi(x, \sigma_2), \qquad \phi(x, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2\sigma^2}. \tag{12}$$

Tukey [1] has used this distribution with $\sigma_2/\sigma_1 = 3$ to demonstrate the sensitivity of the standard deviation to small contaminations. Choosing larger values for the ratio, σ_2/σ_1, mimics the effect of a slowly varying background underneath a sharp peak.

With $f(x)$ given by eq. (12), the definite integrals required to evaluate eq. (11) can be explicitly expressed in terms of elementary functions. These integrals are listed below:

$$\Phi(x, \sigma) \equiv \int_{-\infty}^{x} \phi(x, \sigma)\,\mathrm{d}x,$$

$$\int_{-\infty}^{\infty} \Phi(x, \sigma_1)\phi(x, \sigma_2)\,\mathrm{d}x = \tfrac{1}{2},$$

$$\int_{-\infty}^{\infty} x\Phi(x,\sigma_1)\phi(x,\sigma_2)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\frac{\sigma_2^2}{\sqrt{\sigma_1^2+\sigma_2^2}},$$

$$\int_{-\infty}^{\infty} x^2\Phi(x,\sigma_1)\phi(x,\sigma_2)\,\mathrm{d}x = \sigma_2^2\int_{-\infty}^{\infty}\Phi(x,\sigma_1)\phi(x,\sigma_2)\,\mathrm{d}x,$$

$$\int_{-\infty}^{\infty} \Phi^2(x,\sigma_1)\phi(x,\sigma_2)\,\mathrm{d}x = \frac{1}{\pi}\tan^{-1}\left(\frac{\sqrt{\sigma_1^2+2\sigma_2^2}}{\sigma_1}\right),$$

$$\int_{-\infty}^{\infty} x^2\Phi^2(x,\sigma_1)\phi(x,\sigma_2)\,\mathrm{d}x = \frac{1}{\pi}\frac{\sigma_1\sigma_2^2}{(\sigma_1^2+\sigma_2^2)\sqrt{\sigma_1^2+2\sigma_2^2}} + \sigma_2^2\int_{-\infty}^{\infty}\Phi^2(x,\sigma_1)\phi(x,\sigma_2)\,\mathrm{d}x,$$

$$\int_{-\infty}^{\infty} \Phi(x,\sigma_1)\Phi(x,\sigma_2)\phi(x,\sigma_2)\,\mathrm{d}x = \frac{1}{4}\left(1+\frac{2}{\pi}\tan^{-1}\left(\frac{\sigma_2}{\sqrt{2\sigma_1^2+\sigma_2^2}}\right)\right), \tag{13}$$

$$\int_{-\infty}^{\infty} x^2\Phi(x,\sigma_1)\Phi(x,\sigma_2)\phi(x,\sigma_2)\,\mathrm{d}x = \frac{1}{2\pi}\left(\frac{1}{2}+\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)\frac{\sigma_2^3}{\sqrt{2\sigma_1^2+\sigma_2^2}}$$
$$+\sigma_2^2\int_{-\infty}^{\infty}\Phi(x,\sigma_1)\Phi(x,\sigma_2)\phi(x,\sigma_2)\,\mathrm{d}x.$$

The expected value for the mean difference is:

$$\langle g\rangle = \frac{2}{\sqrt{\pi}}\left[(1-\epsilon)^2\sigma_1 + \epsilon(1-\epsilon)\sqrt{2(\sigma_1^2+\sigma_2^2)} + \epsilon^2\sigma_2\right]. \tag{14}$$

Unlike the mean deviation, the mean difference is not a strictly linear function of the mixing parameter, ϵ. The complete analytic expression for $\langle D^2\rangle$ is too expansive to be included here.

To compare the efficiency of various statistics it is convenient to define a dimensionless parameter which reflects the relative magnitude of the variance. For any statistic, s:

$$\eta(s) \equiv \frac{\lim_{n\to\infty} n\ \mathrm{var}(s)}{\langle s\rangle^2}. \tag{15}$$

With this definition, the asymptotic relative efficiency, ARE, of statistic s_1 relative to s_2 is given by the ratio:

$$\mathrm{ARE} = \eta(s_2)/\eta(s_1). \tag{16}$$

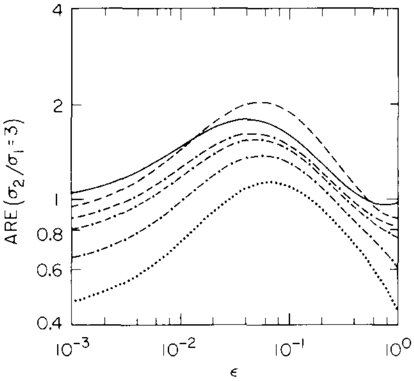

Fig. 1. The asymptotic relative efficiency of the mean deviation and mean difference as a function of the mixing parameter, ϵ, for $\sigma_2/\sigma_1 = 3$. The various curves are plotted with $g_0 = \cdots$, $g_1 = \cdot - \cdot -$, $g_2 = \cdot - - \cdot -$, $g_3 = \cdot - - - \cdot$, $g =$ ———, and $d = - - -$.

Following the steps given above, the ARE was computed for both the mean deviation and mean difference relative to the standard deviation. The results are plotted as a function of the mixing parameter, ϵ, in fig. 1 for $\sigma_1/\sigma_2 = 3$ and in fig. 2 for $\sigma_1/\sigma_2 = 10$. From these graphs two conclusions can be drawn. First of all, the behavior of the mean deviation and mean difference are quite similar although the mean deviation is generally somewhat better. Secondly, the performance of the mean differences based on restricted subsets of pairs is not drastically worse than the statistic computed from all possible pairs. This result demonstrates that the restricted sum variants of the mean difference are reasonably efficient as statistical measures as well as relatively simple to compute.

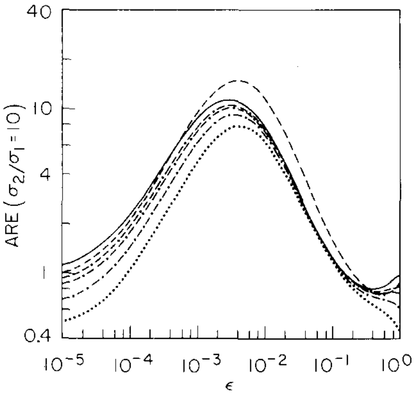

Fig. 2. The asymptotic relative efficiency of the mean deviation and mean difference as a function of the mixing parameter, ϵ, for $\sigma_2/\sigma_1 = 10$. The various curves are plotted with $g_0 = \ldots$, $g_1 = \cdot - \cdot - \cdot$, $g_2 = \cdot - - \cdot -$, $g_3 = \cdot - - - \cdot$, $g =$ ———, and $d = - - -$.

The ARE is a measure of the stability of a given statistic for a well defined distribution. As the mixing parameter rises toward unity the statistic becomes heavily weighted towards its value for the broader distribution. From a practical point of view the statistic no longer usefully measures the width of the narrower distribution which is generally the one of interest. Furthermore the broader background may be subject to large fluctuations in time and so the utility of the statistical measure becomes nil. A simple way of estimating the tolerable level of background distribution is to calculate the mixing parameter required to increase the value of a given statistic by a factor of 2. These results are shown in fig. 3 for the standard deviation, mean deviation, and mean difference as a function of σ_2/σ_1. Because of their essentially linear nature, the mean deviation and mean difference are substantially more tolerant of contamination.

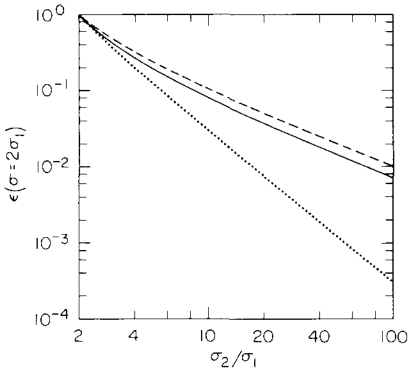

Fig. 3. The mixing parameter, ϵ, required to double the standard deviation (*dotted* line), mean deviation (*dashed* line), and mean difference (*solid* line) as a function of σ_2/σ_1.

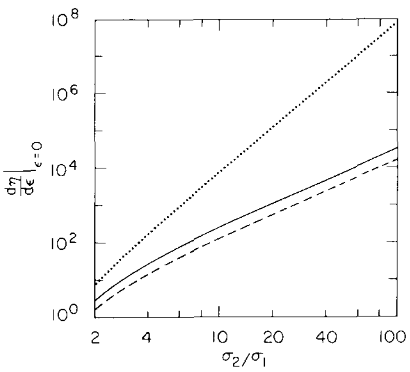

Fig. 4. $d\eta/d\epsilon$ at $\epsilon = 0$ as a function of σ_2/σ_1 for the standard deviation (*dotted* line), mean deviation (*dashed* line), and mean difference (*solid* line).

Another way of looking at the effect of small admixtures of contamination is to calculate the derivative of η with respect to ϵ. In order that a statistic be reasonably accurately determined, $\epsilon\, d\eta/d\epsilon$ must be much less than the sample size, n. In fig. 4, $d\eta/d\epsilon$ at $\epsilon = 0$ is plotted as a function of σ_2/σ_1 for the same three statistics as above. On the scale used, all of the curves for g_i lie essentially on top of each other. Again we conclude that the mean deviation and mean difference will yield significantly better results for a limited sample size.

In the limit of a pure Gaussian distribution, $\langle g \rangle$ is related to the standard deviation by:

$$\sigma = \frac{\sqrt{\pi}}{2}\langle g \rangle. \tag{17}$$

A useful property of the mean deviation is the relative constancy of the ratio, $\sigma/\langle g \rangle$, for other tightly bunched distributions. For a rectangular distribution, $\sigma/\langle g \rangle = \frac{1}{2}\sqrt{3}$ and for a triangular distribution, $\sigma/\langle g \rangle = \frac{1}{2}\sqrt{150/49}$. By comparison, the corresponding ratios for the mean deviation are $\sigma/\langle d \rangle = \sqrt{4/3}$, $\sqrt{3/2}$, and $\sqrt{\pi/2}$ for the rectangular, triangular, and normal distributions respectively.

In addition to eqs. (2) and (3) there are several other equivalent forms of the mean difference. One of these was proposed by Downton [8] and Barnett et al. [9] as an order statistic; the equivalence to Gini's mean difference was noticed by David [5]. Assume that the sample data are ordered by value:

$$x_{(1)} \leqslant \dots \leqslant x_{(i-1)} \leqslant x_{(i)} \leqslant x_{(i+1)} \leqslant \dots \leqslant x_{(n)}.$$

Then

$$g = \frac{2}{n(n-1)} \sum_{i=1}^{n} (2i - n - 1) x_{(i)}. \tag{18}$$

From a computational point of view this is a more efficient expression than eq. (2) since, in the limit of large n, the number of arithmetic operations required grows as $n \log n$ rather than n^2. Another set of equivalent representations for g is:

$$g = \frac{2}{n}\left[\sum_{1 \leqslant i \leqslant n}^{n} x_i - \frac{2}{n-1} \sum_{\substack{1 < i \leqslant n \\ 1 \leqslant j < i}} \min(x_i, x_j)\right] = \frac{2}{n}\left[\frac{2}{n-1} \sum_{\substack{1 < i \leqslant n \\ 1 \leqslant j < 1}} \max(x_i, x_j) - \sum_{1 \leqslant i \leqslant n} x_i\right], \tag{19}$$

which follows from the identity:

$$\frac{x+y}{2} - \min(x, y) = \tfrac{1}{2}|x - y| = \max(x, y) - \frac{x+y}{2}. \tag{20}$$

This form of the mean difference extends in an obvious fashion to the restricted sum variants given by eq.

(3). From some purposes it may be more convenient to average the min and max functions as rough indicators of the active range of the data. If the sample data is grouped as in a histogram, then g is most easily calculated from the sums or integrals implied by eq. (10) using for $f(x)$ the empirical distribution. Note that for such a discrete distribution the appropriate sums can be obtained in a single pass.

In conclusion, it has been shown that the mean difference has a number of useful properties which are ideal for determining the distribution width of approximately Gaussian experimental data. We have seen that, for practical applications, the mean difference is a considerably more reliable statistical measure than the standard deviation although not quite as statistically efficient as the mean deviation. For monitoring real time data streams in which the sample mean is not immediately available and the sample size is large, the restricted sum variants of the mean difference are particularly appropriate. Finally, some of the calculations presented here should provide a warning to experimenters who blithely assume the inevitable suitability of least-square fitting procedures.

The author gratefully acknowledges the help and encouragement of L. Moses. This work was supported by the U.S. Department of Energy.

References

[1] J.W. Tukey, Contributions to probability and statistics, essays in honor of Harold Hotelling, ed., I. Olkin et al., (Stanford University Press, Stanford, California, 1960) p. 448.

[2] R.A. Fisher, Monthly Notices Roy. Astron. 80 (1920) 758.

[3] C. Gini, Studi Economico-giuridic della R. Università di Cagliari (1912).

[4] F.R. Helmert, Astron. Nach. 88 (1976) 127.

[5] For a review of earlier work on the mean difference, see H.A. David, Biometrika 55 (1968) 573.

[6] Z.A. Lomnicki, Ann. Math. Statist. 23 (1952) 635.

[7] M. Kendall and A. Stuart, The advanced theory of statistics, vol. 1, 4th ed. (Macmillan, New York, 1977).

[8] F. Downton, Biometrika 53 (1966) 129.

[9] F.C. Barnett, K. Mullen and J.G. Saw, Biometrika 54 (1967) 551.