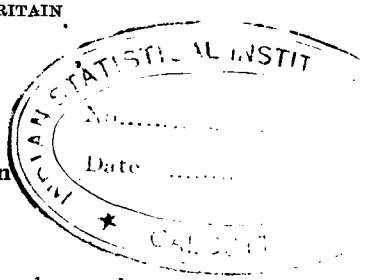
169/4

[*From* Biometrika, *Vol. XXXV. Parts III and IV, December* 1948.]

PRINTED IN GREAT BRITAIN



Note on the median of a multivariate distribution

BY J. B. S. HALDANE

The median of a univariate distribution is an exceedingly useful parameter but, whereas the notions of the mean and mode can be applied without ambiguity to distributions in two or more dimensions, this is not so for the median. It is the object of this note to point out that when we are dealing with multivariate distributions, there are two quite distinct sets of parameters, each of which possess some of the properties of the univariate median, while lacking others.

The possibility arises from the fact that the univariate median is a location parameter associated with two quite different scale parameters. In the first place, for the distribution $dF = f(x)\,dx$, the median is defined as M, where

$$\int_{-\infty}^{M} dF = \int_{M}^{\infty} dF = \tfrac{1}{2}.$$

Integration here and throughout is understood in Stieltjes's sense.

When so defined, the median is obviously associated with the quartiles defined by $\int_{-\infty}^{Q_1} dF = \frac{1}{4}$ and $\int_{Q_3}^{\infty} dF = \frac{1}{4}$, and with the interquartile range $Q_3 - Q_1$.

Secondly, however, we may define the median as the value M which minimizes $\int_{-\infty}^{\infty} |M-x|\,dF$.

Similarly, the mean can be defined as the value m which minimizes $\int_{-\infty}^{\infty} (m-x)^2\,dF$. Now the *minimum* value of this quantity is simply the variance. Just as the mean is associated with the standard deviation as a measure of dispersion, so on this definition the median is associated with the mean deviation about the median. The more commonly used measure, the mean deviation about the mean, has perhaps less to recommend it, since it is not a stationary value, and therefore more liable to error if the corresponding scale parameter is in error. In geometrical language the median is the point the sum of whose distances from the representative points of the sample is a minimum.

Both these definitions of the median are equivalent in the univariate case, and both are of course indeterminate if the number of members of a sample is even, unless an even number of them coincide with the median. The various devices which avoid this indeterminacy represent the median as a limit.

When we pass to two or more dimensions these two definitions are no longer equivalent, and it seems worth while to distinguish the two analogues of the univariate median as the arithmetic and geometric medians.

If we have a number of variates $x, y, z, \ldots$ the arithmetic median is the set of values $(X, Y, Z, \ldots)$, where $X, Y, Z, \ldots$ are the medians of $x, y, z, \ldots$ defined in either of the two above ways. When x, y, z, etc., are different in kind it is obviously the only reasonable generalization. It has the merit of being invariant, like the median, when any of the variates is replaced by a monotonic function of it. But it is not invariant under a rotation of axes.

For consider the arithmetical median of three coplanar points. If we take rectangular axes their co-ordinates are (x_1, y_1), (x_2, y_2), (x_3, y_3). Those of the median are (x_m, y_m), where x_m is the middle value of x_1, x_2, x_3 if they are all different, and the value of the two equal ones if two are equal. Hence as we rotate the axes we find that the position of the arithmetic median changes. Unless it coincides with one of the apices of the triangle, one of the sides must subtend a right angle at it. In fact, the locus consists of those arcs of the circles which have the sides of the triangle as diameters which lie within the triangle (see Fig. 1). If the triangle has a right or obtuse angle, this angle lies on the locus, as does the foot of the perpendicular from it on the opposite side. If the triangle is acute angled, it passes through the feet of the three perpendiculars. Similarly, the locus of the arithmetic median of the vertices of a tetrahedron consists of portions of spheres. For an odd number of more than three points in a plane, the arithmetic median may always be one of them, or its locus may consist of a series of circular arcs. For an even number it is of course indeterminate, unless one or other of the special conventions devised for the univariate case is used.

The geometric median is defined as the point such that the sum of its distances from the sample points is a minimum. It is invariant under a change of axes, but is not invariant when the scales in different directions are altered. Its sole value is therefore in problems of geometrical probability. It occurred in a problem of this type during the recent war, and might perhaps be of value in studies on such aggregates as star clusters, where we desire to find a representative point which is less affected than the centroid by outliers which may not be members of the cluster.

The geometrical median of three coplanar points is the point in the triangle formed by them at which each side subtends an angle of $\frac{2}{3}\pi$, provided that no angle of the triangle exceeds $\frac{2}{3}\pi$. If one angle exceeds this value, the geometrical median is the obtuse vertex of the triangle. I have been unable to find any simple geometrical construction in the case of more than three points. It is, however, easy to show that

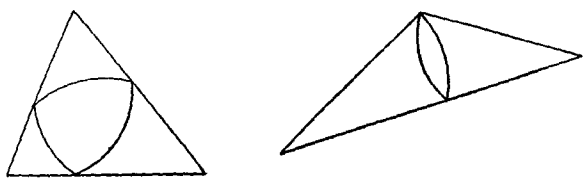


Fig. 1.

the geometrical median is unambiguously defined. For let us take it (or *per impossible*, one of the geometrical medians) as our origin of Cartesian co-ordinates. Consider a set of n coplanar points (x_r, y_r). First suppose that no sample point coincides with the origin, and if necessary rotate the axes so that no axis passes through a sample point. Let R be the sum of the distances of the sample points from the point $(x, 0)$. Then

$$R = \sum_{r=1}^{n} [(x-x_r)^2+y_r^2]^{\frac{1}{2}}, \quad \frac{dR}{dx} = \sum_{r=1}^{n} [(x-x_r)\{(x-x_r)^2+y_r^2\}^{-\frac{1}{2}}].$$

This must be zero when $x = 0$. But

$$\frac{d^2R}{dx^2} = \sum_{r=1}^{n} [y_r^2\{(x-x_r)^2+y_r^2\}^{-\frac{3}{2}}].$$

All the terms in this sum are necessarily positive, since the denominator is the cube of a distance which is taken as positive and can never change its sign. Hence d^2R/dx^2 is always positive, and R has only one minimum.

Next suppose that the median coincides with one of the points, say the first; then

$$R = |x| + \sum_{r=2}^{n} [(x-x_r)^2+y_r^2]^{\frac{1}{2}}, \quad \frac{dR}{dx} = \pm 1 + \sum_{r=2}^{n} [(x-x_r)\{(x-x_r)^2+y_r^2\}^{-\frac{1}{2}}],$$

d^2R/dx^2 is positive as before, but dR/dx has a saltus at $x = 0$, increasing in value by 2, and changing sign. R has therefore a sharp minimum. The proof in three or more dimensions is analogous. Changing to polar co-ordinates with the origin as centre and the co-ordinates of the sample points as (ρ_r, θ_r), it follows that if the median does not coincide with one of them,

$$\Sigma \cos\theta_r = \Sigma \sin\theta_r = 0,$$

whilst if it does so, these sums lie between ± 1. If several sample points coincide with the geometric median, the modifications are obvious.

It is clear that the minimum sum of the distances, divided by the number of points, is the many-dimensional analogue of the mean deviation from the median in one dimension.

To sum up, the arithmetical median is obviously to be preferred in ordinary statistical work, but the geometrical median has certain advantages in problems of geometrical probability. In either case it is desirable to state clearly how the median is defined.