THE EFFICIENCY OF THE SAMPLE MEDIAN

by

Harold H otelling

Research under contract NR 042 031 with the Office of Naval Research for research in probability and statistics at Chapel Hill. Presented September 11, 1954 at the Montreal meeting of the Institute of Mathematical Statistics.

Institute of Statistics
Mimeograph Series No. 117
September, 1954

THE EFFICIENCY OF THE SAMPLE MEDIAN[1]

by Harold Hotelling
Institute of Statistics, University of North Carolina

A random sample of 2n + 1 is drawn from a univariate distribution concerning which nothing is known excepting that its variance is finite. What, if any, least upper bound exists for the ratio of the mean square error in the sample median, regarded as an estimate of the population median, to the mean square deviation in the population from the population median? We answer: This ratio can be unity but no more. For symmetric populations this means that the sampling variance of the median is less than or equal to the variance of the population. If the population variance is infinite, that of the sample median may be finite or infinite. Thus for the Cauchy distribution the variance of the sample median is infinite for samples of three but finite for larger samples. Our theorem deals only with populations of finite variance.

If the population has only two values, each with probability 1/2, the distribution of the median of any random sample of odd size is exactly the same, and its variance therefore the same, as for a single observation from the same population. We shall show that this is the extreme case. But the existence of this case, in which the median fails to gain in accuracy as the sample size increases, is a warning against the use of the median in the absence of any knowledge of the form of the basic distribution, and eliminates an argument occasionally given for using the median as a statistic of location rather than the arithmetic mean. It is true that the arithmetic mean also fails for certain distributions, typified by that of Cauchy, to gain in accuracy with the size of the sample; but the Cauchy distribution with its infinite high tails is much less like those to be expected in the run of applications than is the symmetrical two-point distribution, which at least has the realistic property of a finite range.

Adverse reflections on the use of the median result not only from this remark that it is sometimes no more accurate than a single observation taken at random, but also from recent work on order statistics. Thus Smith and Jones [1], whose work was further extended by Siddiqui [2], find that in samples from the Laplace double exponential distribution,

---

1 Research under contract NR 042 031 with the Office of Naval Research for research in probability and statistics at Chapel Hill. Presented September 11, 1954 at the Montreal meeting of the Institute of Mathematical Statistics. This study grew out of a question raised by Mr. Darrell Bordelon of the Naval Ordnance Laboratory.

$$\frac{1}{2} e^{-|x-\mu|} dx,$$

for which the median is the maximum likelihood estimate of μ, certain linear functions of ordered observations have smaller variances and are unbiased. For random samples of three, $x_1 \leq x_2 \leq x_3$, they obtain as the unbiased linear estimate of minimum variance.

$$(4x_1 + 19x_2 + 4x_3)/27.$$

I calculate from their covariances among the ordered observations that this linear function has variance .5895, whereas the median x_2 has variance .6381. This is a notable case of the maximum likelihood estimate failing to give minimum variance among unbiased estimates of the same parameter. However the advantage of about eight per cent for samples of three fades to zero with increase in size of the sample, while the linear estimates have increasingly complicated sets of coefficients; so for samples of sizes above the smallest it appears reasonable to continue to use the median.

The moral seems to be, not that the median is a bad statistic, but that it should not be used indiscriminately in the absence of any information about the form of the population sampled. When it is used, probabilities relevant to the inferences to be made should be calculated from its distribution, exact or approximate. These calculations, particularly approximations for large samples, may be facilitated by the results of [3].

We shall now prove that the mean square error in the sample median cannot exceed that of a random observation, - a proposition which may at first sight appear self-evident.

Without loss of generality we shall assume the population median of the variate x to be zero, and also that $\int x^2 = 1$. With the notation $\underline{m}$ for the median value of x in a sample of $2n + 1$ independent values of x, our theorem will be proved by showing that

$$\int \underline{m}^2 \leq 1.$$

Denoting by $F(x)$ the cumulative distribution function of the population, we shall use the inverse of this function, $x = x(F)$, making this inverse definite on the open segment $0 < F < 1$ by putting $x(\frac{1}{2}) = 0$ in accordance with the assumption of

zero population median, and otherwise assigning to $x(F)$ the minimum value of x for which $F(x)$ takes a specified value. Then $x(F)$ is non-decreasing.

The distribution of the median is well known to be given by

$$C \left[F(m)\right]^n \left[1 - F(m)\right]^n \, dF(m),$$

where $C = (2n+1)\,!\,(n\,!)^{-2}$. Let us put

$$g(F) = C\, F^n (1-F)^n.$$

We thus have

$$\xi m^2 = \int_0^1 \left[x(F)\right]^2 \, g(F) \, dF.$$

It is convenient to use the function

$$u(F) = \left[x(F)\right]^2 + \left[x(1-F)\right]^2 - 2.$$

The non-decreasing character of $x(F)$ and the fact that $x(F) \leq 0$ for $F < \frac{1}{2}$ and ≥ 0 for $F > \frac{1}{2}$ make it clear that $u(F)$ is non-increasing for $F < \frac{1}{2}$ and non-decreasing for $F > \frac{1}{2}$. From the definition of $g(F)$,

$$g(F) = g(1 - F), \qquad \int_0^1 g(F) \, dF = 1.$$

Since we have taken ξx^2 to be unity,

$$1 = \int_0^1 \left[x(F)\right]^2 dF = \int_0^{\frac{1}{2}} \left\{ \left[x(F)\right]^2 + \left[x(1-F)\right]^2 \right\} dF = \int_0^{\frac{1}{2}} u(F) \, dF + 1,$$

so that

$$\int_0^{\frac{1}{2}} u(F) \, dF = 0.$$

From these relations we have

$$\mathcal{E}\, m^2 = \int_0^{\frac{1}{2}} \{ [x(F)]^2 + [x(1-F)]^2 \}\ g(F)\, dF$$

$$= \int_0^{\frac{1}{2}} \{u(F) + 2\}\ g(F)\, dF$$

$$= \int_0^{\frac{1}{2}} u(F)\ g(F)\, dF + 1.$$

Since $u(F)$ equals -2 for $F = \frac{1}{2}$ and is non-increasing on $0 < F < \frac{1}{2}$, and since its integral is zero, there exists a number F_o, $0 < F_o < \frac{1}{2}$, such that $u(F) \geq 0$ for $0 < F < F_o$ and $u(F) < 0$ for $F_o < F < \frac{1}{2}$. Furthermore, since $g(F)$ is an increasing function on $0 \leq F \leq \frac{1}{2}$, its values are less than $g(F_o)$ for $F < F_o$ and greater than $g(F_o)$ for $F_o < F < \frac{1}{2}$. Consequently,

$$\int_0^{F_o} u(F)\ g(F)\, dF \leq g(F_o) \int_0^{F_o} u(F)\, dF,$$

and

$$\int_{F_o}^{\frac{1}{2}} |u(F)|\ g(F)\, dF \geq g(F_o) \int_{F_o}^{\frac{1}{2}} |u(F)|\, dF.$$

In the last integral $u(F) \leq 0$ so that the inequality is reversed when the absolute value signs are removed. Combining with the preceding inequality gives

$$\int_0^{\frac{1}{2}} u(F)\ g(F)\, dF \leq g(F_o) \int_0^{\frac{1}{2}} u(F)\, dF = 0$$

Therefore

$$\mathcal{E}\, m^2 \leq 1,$$

as was to be proved.

BIBLIOGRAPHY

[1] John H. Smith and Howard L. Jones, The weighted mean of random observations arranged in order of size. National Bureau of Standards Statistical Engineering Laboratory. February 1951. Hectographed.

[2] M. M. Siddiqui, The estimation of parameters of a double exponential universe by ordered observations. Unpublished M. A. thesis, American University, Washington, D. C. 1954.

[3] John T. Chu and Harold Hotelling, "The moments of the sample median." Presented concurrently with this paper.