ON THE DISTRIBUTION OF THE SAMPLE MEDIAN[1]

By John T. Chu

University of North Carolina

1. Summary. Upper and lower bounds are obtained for the cumulative distribution function of the sample median of a sample of size $2n + 1$ drawn from a continuous population. It is shown that if the parent population is normal, then the distribution of the sample median tends "rapidly" to normality. Other kinds of parent populations are also discussed.

2. Introduction. Let a continuous population be given with cdf $F(x)$ (cumulative distribution function) and median ξ (assumed to exist uniquely). For a sample of size $2n + 1$, let $\tilde{x}$ denote the sample median. The distribution of $\tilde{x}$, under certain conditions, is known ([2], p. 369) to be asymptotically normal with mean ξ and variance $\sigma_n^2 = 1/4\,[f(\xi)]^2(2n + 1)$, where $f(x) = F'(x)$ is the pdf (probability density function).

Several authors, among them Hojo [4] and Cadwell [1], have stated that numerical investigations showed that, if the parent population is normal, "the convergence (of the distribution of $\tilde{x}$) to normality is surprisingly fast." However, no mathematical proof or disproof seems ever to have been given for this experimental result.

This paper shows mathematically that the findings are correct. Upper and lower bounds are obtained for $P[-x < (\tilde{x} - \xi) / \sigma_n < y]$ in (8) and (9). If no very high accuracy is required, these bounds are reduced to a simpler form in (10) and (11). Examination of these bounds makes it evident that the distribution of $(\tilde{x} - \xi) / \sigma_n$ tends "rapidly" to normality.

Rectangular and Laplace parent populations are briefly discussed. It seems that in these cases the distribution of $(\tilde{x} - \xi) / \sigma_n$ tends to normality at a "much slower speed."

3. Upper and lower bounds. Let $F(x)$ and $f(x)$ be respectively the cdf and pdf of a certain population whose median is ξ. If $g(x)$ is the pdf of the sample median $\tilde{x}$ of a sample of size $2n + 1$, then

$$g(x) = C_n[F(x)]^n[1 - F(x)]^n f(x), \qquad C_n = (2n + 1)! / n!n!.$$

If $f(\xi) \neq 0$ and $f'(x)$ is continuous in some neighborhood of $x = \xi$, then $\tilde{x}$ is known ([2], p. 369) to have an asymptotically normal distribution with mean ξ and variance

$$\sigma_n^2 = \frac{1}{4[f(\xi)]^2\,(2n + 1)}.$$

Received June 11, 1954.

[1] Special report to the Office of Naval Research of work at Chapel Hill under Contract N7-onr-28402 for research in probability and statistics.

For finite n, let the cdf of $(\tilde{x} - \xi) / \sigma_n$ be $H(x)$. Then for any $y > 0$,

$$H(y) - \tfrac{1}{2} = \int_{\xi}^{\xi+y\sigma_n} g(t)\,dt = C_n \int_{1/2}^{F(\xi+y\sigma_n)} u^n (1-u)^n\,du$$

$$= (\tfrac{1}{2})^{2n} C_n \int_0^{F(\xi+y\sigma_n)-1/2} (1 - 4v^2)^n\,dv.$$

Applying the transformations

$$v = \begin{cases} av', & 0 < a \leqq 1; \\ bv', & 1 \leqq b; \end{cases} \qquad v' = \tfrac{1}{2}\sqrt{1 - \exp\left[-t^2/(2n+1)\right]},$$

we obtain without difficulty

$$(1)\quad H(y) - \tfrac{1}{2} \geqq aB_n \int_0^{t_1} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{n+1}{2n+1}\,t^2\right] h_1\left(\frac{t}{\sqrt{2n+1}}\right) dt, \qquad 0 < a \leqq 1,$$

$$(2)\quad H(y) - \tfrac{1}{2} \leqq bB_n \int_0^{t_2} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{n}{2n+1}\,t^2\right] h_2\left(\frac{t}{\sqrt{2n+1}}\right) dt, \qquad b \geqq 1,$$

where $B_n = (\tfrac{1}{2})^{2n+1} C_n \sqrt{2\pi} \,/\, \sqrt{2n+1}$ and

$$(3)\quad h_1(t) = t \,/\, \sqrt{1 - \exp(-t^2)}, \qquad h_2(t) = t \exp(-t^2) \,/\, \sqrt{1 - \exp(-t^2)};$$

$$(4)\quad \begin{aligned} t_1 &= \sqrt{-(2n+1)\log\{1 - (4/a^2)\,[F(\xi + y\sigma_n) - \tfrac{1}{2}]^2\}}, \\ t_2 &= \sqrt{-(2n+1)\log\{1 - (4/b^2)\,[F(\xi + y\sigma_n) - \tfrac{1}{2}]^2\}}. \end{aligned}$$

It can be shown (by differentiations) that $h_1(t)$ and $h_2(t)$ are respectively monotonically increasing and decreasing functions of t, when $t \geqq 0$, and that $\lim_{t\to 0} h_1(t) = \lim_{t\to 0} h_2(t) = 1$. Hence we obtain from (1) and (2)

$$H(y) - \tfrac{1}{2} \geqq aB_n \sqrt{1 - \frac{1}{2n+2}} \left[\phi\left(t_1 \sqrt{\frac{2n+2}{2n+1}}\right) - \tfrac{1}{2}\right],$$

$$H(y) - \tfrac{1}{2} \leqq bB_n \sqrt{1 + \frac{1}{2n}} \left[\phi\left(t_2 \sqrt{\frac{2n}{2n+1}}\right) - \tfrac{1}{2}\right],$$

where

$$(5)\qquad \phi(t) = \int_{-\infty}^{t} (1/\sqrt{2\pi}) \exp(-\tfrac{1}{2}x^2)\,dx.$$

In a similar way we can show that, for arbitrary $x > 0$ and $y > 0$,

$$(6)\quad \begin{aligned} &H(y) - H(-x) \\ &\quad \geqq aB_n \sqrt{1 - \frac{1}{2n+2}} \left[\phi\left(t_1 \sqrt{\frac{2n+2}{2n+1}}\right) - \phi\left(-t_3 \sqrt{\frac{2n+2}{2n+1}}\right)\right], \\ &H(y) - H(-x) \\ &\quad \leqq bB_n \sqrt{1 + \frac{1}{2n}} \left[\phi\left(t_2 \sqrt{\frac{2n}{2n+1}}\right) - \phi\left(-t_4 \sqrt{\frac{2n}{2n+1}}\right)\right], \end{aligned}$$

where t_3 and t_4 are obtained from t_1 and t_2 in (4) by replacing y by $-x$.

It can be seen from (4) that if x and y are fixed, $t_1 = t_2 = y + O(1)$ and $t_3 = t_4 = x + O(1)$ for large n. The following sections will show, for various kinds of parent distributions, that if a and b are properly chosen, then (6) remains valid if t_1, t_2 and t_3, t_4 are replaced by y and x, respectively.

If n is large, upper and lower bounds for B_n can be obtained by using Stirling's formula. Feller [3] showed that for $n \geqq 4$,

$$n! = \sqrt{2\pi}\, n^{n+1/2} \exp\left[-n + \frac{1}{12n} - \frac{(1+\theta)}{360n^3}\right], \qquad |\theta| \leqq \tfrac{1}{6}.$$

If the last term, $-(1 + \theta) / 360n^3$, is omitted, then it can be shown that

$$1 + \frac{1}{8n} - \frac{7n+3}{24n^2(2n+1)} < B_n < 1 + \frac{1}{8n} + \frac{1}{16n(8n-1)},$$

or $B_n \sim 1 + 1/8n$.

4. Normal parent population. Suppose that a sample of size $2n + 1$ is drawn from a normal population with mean ξ and variance σ^2. The distribution of $\tilde{x}$ is then asymptotically normal with mean ξ and variance $\pi\sigma^2 / 2(2n + 1)$. It has been shown that if, for $x > 0$,

$$(7)\qquad \phi(x) - \phi(-x) = a(x) \sqrt{1 - \exp\left[-(2/\pi)\, x^2\right]},$$

then $a(x)$, a function of x, never exceeds 1 and is very close to 1 for all values of $x > 0$. Williams [6] proved that $a(x) \leqq 1$ and tabulated $1/a(x) - 1$ for a number of values of x ranging from .1 to 2.0. Pólya [5] gave several proofs for the same inequality and remarked that if $\sqrt{1 - \exp\left[-(2/\pi)x^2\right]}$ is used as an approximation to $\phi(x) - \phi(-x)$, "then the error committed is less than one per cent (even less than .71 per cent) of the quantity approximated." In other words, $a(x) > .9929$ for all $x > 0$.

For arbitrary $x > 0$ and $y > 0$, let

$$x_n = \sqrt{\pi/2}\;\; x / \sqrt{2n+1}, \qquad y_n = \sqrt{\pi/2}\;\; y / \sqrt{2n+1}.$$

Applying (7) to (6) yields

$$H(y) - H(-x) \geqq \min\{a(x_n), a(y_n)\} \cdot B_n \sqrt{1 - \frac{1}{2n+2}} \left[\phi\left(y\sqrt{\frac{2n+2}{2n+1}}\right) - \phi\left(-x\sqrt{\frac{2n+2}{2n+1}}\right)\right], \tag{8}$$

$$H(y) - H(-x) \leqq B_n \sqrt{1 + \frac{1}{2n}} \left[\phi\left(y\sqrt{\frac{2n}{2n+1}}\right) - \phi\left(-x\sqrt{\frac{2n}{2n+1}}\right)\right], \tag{9}$$

where $\phi(x)$ and $a(x)$ are defined by (5) and (7). If no very high accuracy is required one may use

$$H(y) - H(-x) \geqq .9929\,(1 + \tfrac{1}{8}n)\sqrt{1 - 1/(2n+2)}\,[\phi(y) - \phi(-x)], \tag{10}$$

$$H(y) - H(-x) \leqq (1 + \tfrac{1}{8}n)\sqrt{1 + 1/2n}\,[\phi(y) - \phi(-x)]. \tag{11}$$

5. Other parent populations.

A. Rectangular distribution. Let $f(x) = 1/(d-c)$, where $c < x < d$. Then $\xi = \frac{1}{2}(c+d)$ and $\sigma_n^2 = (c-d)^2/4(2n+1)$. Let $H(x)$ be the cdf of $(\tilde{x} - \xi)/\sigma_n$. If $x > 0$ and $y > 0$, then lower bounds for $H(y) - H(-x)$ are the right sides of (8) and (10), without the factors min $\{a(x_n), a(y_n)\}$ and .9929. The upper bounds for $H(y) - H(-x)$ are the right sides of (9) and (11) with an additional factor max $\{b(x_n), b(y_n)\}$, where $b(x)$ is defined by

$$b(x) = \sqrt{\frac{x}{1 - e^{-x}}}, \quad x > 0; \qquad x_n = \frac{x^2}{2n+1}; \qquad y_n = \frac{y^2}{2n+1}.$$

We note that $b(x)$ is close to 1 only if x is close to 0, for example, $b(.1) = 1.02$ and $b(.2) = 1.05$. This means that unless x and y are small, upper and lower bounds for $H(y) - H(-x)$ are not very close to each other except for large n.

B. Laplace distribution. Let $f(x) = (1/2\lambda)\exp[-|x-\xi|/\lambda]$, where $-\infty < x < \infty$. Then ξ is the median and $\sigma_n^2 = \lambda^2/(2n+1)$. Define $c(x)$ by

$$1 - e^{-x} = c(x)\sqrt{1 - e^{-x^2}}, \qquad x > 0. \tag{12}$$

We say that $c(x) \leqq 1$. If $x \geqq 1$, this is obvious; if $x \leqq 1$, we use (7) to prove it. It then becomes clear that upper bounds for $H(y) - H(-x)$ are the same as those corresponding to a normal parent population, that is (9) and (11). Lower bounds for $H(y) - H(-x)$ are the right sides of (8) and (10) with min $\{a(x_n), a(y_n)\}$ and .9929 replaced by min $\{c(x_n), c(y_n)\}$, where $c(x)$ is defined by (12) and $x_n = x/\sqrt{2n+1}$. Again we remark that $c(x)$ is close to 1 if x is small or large, for example, $c(.1) = .99$, $c(.2) = .92$, and $c(3) = .97$, while $c(1) = .79$ and $c(2) = .87$.

Finally we note that since $c(x)$ tends to 1 as x tends to 0, $H(y) - H(-x)$ tends to $\phi(y) - \phi(-x)$ as n tends to infinity. Therefore $(\tilde{x} - \xi)/\sigma_n$ has an asymptotically normal distribution. In the general theorem which showed that $(\tilde{x} - \xi)/\sigma_n$

has an asymptotically normal distribution (see] 2], p. 369, also the beginning of Sec. 3), it is required that $f'(\xi)$ be continuous. For a Laplace distribution, however, $f'(\xi)$ does not exist.

REFERENCES

[1] J. H. CADWELL, "The distribution of quantiles of small samples," *Biometrika*, Vol. 39 (1952), pp. 207–211.

[2] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[3] W. FELLER, "On the normal approximation to the binomial distribution," *Ann. Math. Stat.*, Vol. 16 (1945), pp. 319–329.

[4] T. HOJO, "Distribution of the median, quartiles, and interquartile distance in samples from a normal population," *Biometrika*, Vol. 23 (1931), pp. 315–360.

[5] G. PÓLYA, "Remarks on computing the probability integral in one and two dimensions," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1949, pp. 63–78.

[6] J. D. WILLIAMS, "An approximation to the probability integral," *Ann. Math. Stat.*, Vol. 17 (1946), pp. 363–365.