ON THE DISTRIBUTION OF THE SAMPLE MEDIAN

by

John T. Chu

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

Institute of Statistics
Mimeograph Series No. 105
May 1954

On the Distribution of the Sample Median[1]

by John T. Chu
Institute of Statistics, 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 (assume it exists uniquely). For a sample of size 2n+1, let $\tilde{x}$ denote the sample median. It is well-known [2, p. 369] that the distribution of $\tilde{x}$ is, under certain conditions, 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 T. Hojo [4] and J. H. Cadwell [1], claimed that numerical investigations showed that, if the parent population is normal, "the convergence (of the distribution of $\tilde{x}$) to normality is surprisingly fast." It seems, however, no mathematical proof or disproof has ever been given to this experimental result. It is the purpose of this paper to show mathematically that their findings are correct. Upper and lower bounds are obtained for $P[-x < (\tilde{x} - \xi)/\sigma_n < y]$ (Equations (23) and (24)), and for large samples, these bounds are reduced to a simpler form, (25) and (26). By examining these bounds, it then becomes 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."

---

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.

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

$$(1) \qquad g(x) = C_n \left[F(x)\right]^n \left[1 - F(x)\right]^n f(x),$$

where $C_n = (2n+1)!/n!n!$. It is known [2, P. 369] that if $f(\xi) \neq 0$ and $f'(x)$ is continuous in some neighborhood of $x = \xi$, then $\tilde{x}$ has an asymptotically normal distribution with mean ξ and variance

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

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

$$(3) \qquad H(y) - \tfrac{1}{2} = \int_{\xi}^{\xi + y\sigma_n} g(t)dt = C_n \int_{\frac{1}{2}}^{F(\xi + y\sigma_n)} u^n(1-u)^n \, du$$

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

Applying to (3) the transformation

$$v = \frac{1}{2}\left[1 - e^{-t^2/(2n+1)}\right]^{\frac{1}{2}}, \tag{4}$$

we obtain without difficulty

$$H(y) - \tfrac{1}{2} \geq a\, B_n \int_0^{t_1} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{n+1}{2n+1}t^2}\, h_1\!\left(\frac{t}{\sqrt{2n+1}}\right) dt, \qquad \text{for } 0 \leq a \leq 1, \tag{5}$$

$$H(y) - \tfrac{1}{2} \leq b\, B_n \int_0^{t_2} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{n}{2n+1}t^2}\, h_2\!\left(\frac{t}{\sqrt{2n+1}}\right) dt, \qquad \text{for } b \geq 1, \tag{6}$$

where

$$B_n = \left(\tfrac{1}{2}\right)^{2n+1} C_n \sqrt{2\pi} / \sqrt{2n+1}, \tag{7}$$

$$h_1(t) = t/(1-e^{-t^2})^{\frac{1}{2}} \tag{8}$$

$$h_2(t) = t\, e^{-t^2}/(1-e^{-t^2})^{\frac{1}{2}}, \tag{9}$$

$$(10)\quad t_1 = \left[-(2n+1) \log \left\{ 1 - (4/a^2) \left[F(\xi + y\sigma_n) - \tfrac{1}{2} \right]^2 \right\} \right]^{\frac{1}{2}},$$

$$(11)\quad t_2 = \left[-(2n+1) \log \left\{ 1 - (4/b^2) \left[F(\xi + y\sigma_n) - \tfrac{1}{2} \right]^2 \right\} \right]^{\frac{1}{2}}$$

It can be shown (by differentiations) that $h_1(t)$ and $h_2(t)$ are respectively monotonically increasing and decreasing functions of t, when $t \geq 0$, and that $\lim_{t \longrightarrow 0} h_1(t) = \lim_{t \longrightarrow 0} h_2(t) = 1$. Hence we obtain, from (5) and (6),

$$(12)\quad H(y) - \tfrac{1}{2} \geq a B_n \sqrt{1 - \frac{1}{2n+2}} \left[\Phi \left(t_1 \sqrt{\frac{2n+2}{2n+1}} \right) - \tfrac{1}{2} \right],$$

$$(13)\quad H(y) - \tfrac{1}{2} \leq b B_n \sqrt{1 + \frac{1}{2n}} \left[\Phi \left(t_2 \sqrt{\frac{2n}{2n+1}} \right) - \tfrac{1}{2} \right],$$

where

$$(14)\qquad \Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx.$$

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

$$(15)\quad H(y) - H(-x) \geq a B_n \sqrt{1 - \frac{1}{2n+2}} \left[\Phi \left(t_1 \sqrt{\frac{2n+2}{2n+1}} \right) - \Phi \left(-t_3 \sqrt{\frac{2n+2}{2n+1}} \right) \right],$$

$$(16)\quad H(y) - H(-x) \leq b B_n \sqrt{1 + \frac{1}{2n}} \left[\Phi \left(t_2 \sqrt{\frac{2n}{2n+1}} \right) - \Phi \left(-t_4 \sqrt{\frac{2n}{2n+1}} \right) \right],$$

where t_3 and t_4 are obtained from t_1 and t_2 in (10) and (11) by replacing y by -x.

It can be seen from (10) and (11) 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. In the following sections we will show, for various kinds of parent distributions, that if a and b are properly chosen, then (15) and (16) remain valid if t_1, t_2; and t_3, t_4 are respectively replaced by y and x.

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

$$(17) \qquad n! = \sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n+\frac{1}{12n}-\frac{(1+\theta)}{360n^3}},$$

where $|\theta| \leq 1/6$. If the last term, $-\frac{1+\theta}{360n^3}$, is omitted, then it can be shown that

$$(18) \qquad 1 + \frac{1}{8n} - \frac{7n+3}{24n^2(2n+1)} < B_n < 1 + \frac{1}{8n} + \frac{1}{16n(8n-1)},$$

or

$$(19) \qquad B_n \sim 1 + \frac{1}{8n}.$$

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

$x > 0$,

$$(20) \quad \Phi(x) - \Phi(-x) = a(x)\left[1 - e^{-\frac{2}{\pi}x^2}\right]^{\frac{1}{2}},$$

then $a(x)$, a function of x, never exceeds 1 and is very close to 1 for all values of $x > 0$. J. D. Williams [6] proved that $a(x) \leq 1$ and tabulated $1/a(x) - 1$ for a number of values of x ranging from .1 to 2.0. G. Pólya [5] gave several proofs for the same inequality and remarked that if

$(1 - e^{-\frac{2}{\pi}x^2})^{\frac{1}{2}}$ 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,

$$(21) \quad a(x) > .9929 \quad , \quad \text{for all } x > 0.$$

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

$$(22) \quad x_n = \sqrt{\frac{\pi}{2}}\,\frac{x}{\sqrt{2n+1}} \quad , \quad y_n = \sqrt{\frac{\pi}{2}}\,\frac{y}{\sqrt{2n+1}}$$

Applying (20) to (15) and (16), it follows that

$$(23) \quad H(y) - H(-x) \geq \text{Min}\{a(x_n), a(y_n)\}\, B_n\sqrt{1 - \frac{1}{2n+2}}\left[\Phi\left(y\sqrt{\frac{2n+2}{2n+1}}\right) - \Phi\left(-x\sqrt{\frac{2n+2}{2n+1}}\right)\right]$$

$$(24) \quad H(y) - H(-x) \leq B_n\sqrt{1 + \frac{1}{2n}}\left[\Phi\left(y\sqrt{\frac{2n}{2n+1}}\right) - \Phi\left(-x\sqrt{\frac{2n}{2n+1}}\right)\right],$$

where B_n, $\Phi(x)$, and $a(x)$ are defined by (7), (14), and (20). For large n we suggest to use

$$(25)\quad H(y) - H(-x) \geq .9929\,(1 + \tfrac{1}{8n})\sqrt{1 - \tfrac{1}{2n+2}}\,[\Phi(y) - \Phi(-x)],$$

$$(26)\quad H(y) - H(-x) \leq (1 + \tfrac{1}{8n})\sqrt{1 + \tfrac{1}{2n}}\,[\Phi(y) - \Phi(-x)].$$

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 RHS (right-hand sides) of (23) and (25), without the factors Min $\{a(x_n), a(y_n)\}$ and .9929 and upper bounds for $H(y) - H(-x)$ are the RHS of (24) and (26) with an additional factor Max $\{b(x_n), b(y_n)\}$ where $b(x)$ is defined by

$$(27)\qquad b(x) = \sqrt{\frac{x}{1-e^{-x}}}\quad , \qquad x > 0,$$

and

$$(28)\qquad x_n = x^2/(2n+1)\quad , \qquad y_n = y^2/(2n+1).$$

We note that $b(x)$ is close to 1 only if x is close to 0, e. g., $b(.1) = 1.02$ and $b(.2) = 1.05$. This means: 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 $f(x) = \frac{1}{2\lambda}\, e^{-\frac{|x-\xi|}{\lambda}}$,

where $-\infty < x < \infty$, then ξ is the median and $\sigma_n^2 = \lambda^2/(2n+1)$. Define $c(x)$ by

(29) $$1 - e^{-x} = c(x)\,(1 - e^{-x^2})^{\frac{1}{2}} \qquad , x > 0.$$

We say that $c(x) \leq 1$: if $x \geq 1$, this is obvious; if $x \leq 1$, use (20) 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, i.e., (24) and (26). Lower bounds for $H(y) - H(-x)$ are the RHS of (23) and (25) 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 (29) and $x_n = x/\sqrt{2n+1}$. Again we remark that $c(x)$ is close to 1 if x is small or large, e. g., $c(.1) = .99$, $c(.2) = .92$, 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 § 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, Berkeley, 1949, pp. 63-78.

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