# DETERMINING SAMPLE SIZE FOR A SPECIFIED WIDTH CONFIDENCE INTERVAL

By Franklin A. Graybill

*Oklahoma State University*

**1. Introduction.** If an experimenter decides to use a confidence interval to locate a parameter, he is concerned with at least two things: (1) Does the interval contain the parameter? (2) How wide is the interval? In general the answer to these questions cannot be given with absolute certainty, but must be given with a probability statement. If we let $\alpha$ be the probability that the interval contains the parameter, and let $\beta^2$ be the probability that the width is less than $d$ units, then the general procedure is to fix $\alpha$ in advance and compute $\beta^2$. The value of $\beta^2$ is in general a function of the positive integer $n$, the sample size by which the confidence interval is computed. ($\beta^2$ is also a function of $\alpha$). In most confidence intervals, $\beta^2$ increases as $n$ increases. For any particular situation $\beta^2$ may be too low to be useful, hence an experimenter may wish to increase $\beta^2$ by taking more observations (increasing $n$). The problem the experimenter then faces is the determination of $n$ such that (A) the probability will be equal to $\alpha$ that the confidence interval contains the parameter, and (B) the probability will be equal to $\beta^2$ that the width of the confidence interval will be less than $d$ units (where $\alpha$, $\beta^2$, and $d$ are specified).

To solve this problem will generally require two things: (1) The form of the frequency function from which the sample of size $n$ is to be selected; (2) Some previous information on the unknown parameters in the frequency function.

This suggests that the sample be taken in two steps; the first sample will be used to determine the number of observations to be taken in the second sample so that (A) and (B) will be satisfied.

For a confidence interval on the mean of a normal population with unknown variance this problem has been solved by Stein [1] for $\beta^2 = 1$.

The purpose of this paper is to determine $n$, to satisfy (A) and (B) for distributions other than the normal.

**2. Theory.** Suppose $X$ is the width of a confidence interval on a parameter $\mu$ with confidence coefficient $\alpha$. Suppose further that it is desired that the probability be $\beta^2$ that $X$ be less than $d$. The problem is to determine $n$, the number of observations, on which to base $X$. Since $n$ depends on the random variables used in step one, $n$ is a random variable.

We will prove the following (we will use the notation $P(A)$ for the probability that the event A occurs):

Theorem. *Let the chance variable $X$ be the width of a confidence interval on a parameter $\mu$ based on a sample of size n. Suppose that $X$ depends on n and on an unknown parameter $\theta$ '$\theta$ may be the parameter $\mu$). Suppose also that there exists a*

Received April 29, 1957; revised October 8, 1957.

*function of $X$, $\theta$, and $n$, say $g(X; \theta, n)$, such that if $Y = g(X; \theta, n)$, then the distribution of $Y$ does not depend on any unknown parameters except $n$. Let $f(n)$ be a function of $n$ such that*

(1) $$P[Y < f(n)] = \beta \quad \text{for any} \quad 0 < \beta < 1.$$

*Let the solution of the equation $g(x; \theta, n), = f(n)$ for $x$ be $x = h(\theta, n)$, and suppose the following are true for $x > 0$:*

(2)
- (a) *$g(x; \theta, n)$ is monotonic increasing in $x$ for every $n$ and $\theta$.*
- (b) *$h(\theta, n)$ is monotonic increasing for every $n$.*
- (c) *$h(\theta, n)$ is monotonic decreasing in $n$ for every $\theta$.*
- (d) *$z$ is random variable which is available from step one of the procedure such that $P[t(z) > \theta] = \beta$ for $0 < \beta < 1$, where $t(z)$ is a function of $z$ which does not depend on any unknown parameters or on $n$.*

*Let $d$ and $\beta$ be specified in advance. Then if $n$ is such that the equation*

(3) $$h[t(z), n] \leqq d$$

*is satisfied ($t(z)$ is known) then the following inequality is true:*

(4) $$P(X \leqq d) \geqq \beta^2.$$

Proof. Substituting into Eq. (1) we get

(5) $$P[g(X; \theta, n) < f(n)] = \beta.$$

Solving for $X$ and using 2(a) gives us

(6) $$P[X < h(\theta, n)] = \beta.$$

For any $\theta_1 \geqq \theta$ we can use 2(b) and obtain

(7) $$P[X < h(\theta_1, n) \mid \theta_1 \geqq \theta] \geqq P[X < h(\theta, n)] = \beta.$$

By considering the joint distribution of $X$ and $t(z)$ we can write

(8) $$P(X \leqq d) \geqq P[X \leqq d, t(z) > \theta] = P[X \leqq d \mid t(z) > \theta] \cdot P[t(z) > \theta].$$

If $n$ is any integer satisfying

(9) $$h[t(z), n] \leqq d,$$

we can use (7), 2(c), and 2(d) in Eq. (8) and obtain

$$P(X \leqq d) \geqq \beta^2.$$

If the function in 2(b) is monotonic *decreasing*, then the theorem is also true but the inequality in 2(d) must be reversed. The theorem is also true if the function in 2(a) is monotonic *decreasing*. The conditions (2) may appear quite stringent; however, many of the functions in common use in statistics satisfy these conditions.

## 3. Illustrations

*Example* 1. Suppose we want an $\alpha$ confidence interval on the variance $\sigma^2$ of a normal population to be less than $d$ units in length with probability of $\beta^2$.

We will define

$$s_n^2 = \frac{1}{n-1}\sum_{1=1}^{n}(v_i - \bar{v})^2,$$

where $v_i$ is distributed normally with mean $\mu$ and variance $\sigma^2$. An $\alpha$ confidence interval on $\sigma^2$ is given by

$$P\left[\frac{(n-1)\,s_n^2}{\chi_1^2(n)} \leqq \sigma^2 \leqq \frac{(n-1)\,s_n^2}{\chi_2^2(n)}\right] = \alpha,$$

where $\chi_1^2(n)$ and $\chi_2^2(n)$ are such that

$$\int_0^{\chi_2^2(n)} W(\chi^2; n)\, d\chi^2 = \frac{1-\alpha}{2},$$

$$\int_{\chi_1^2(n)}^{\infty} W(\chi^2; n)\, d\chi^2 = \frac{1-\alpha}{2},$$

where $W(\chi^2; n)$ is a Chi-square frequency function with $n - 1$ degrees of freedom. The width of the interval is

$$X = (n-1)\,s_n^2\left[\frac{1}{\chi_2^2(n)} - \frac{1}{\chi_1^2(n)}\right].$$

If we let

$$\frac{1}{\chi_2^2(n)} - \frac{1}{\chi_1^2(n)} = C_n,$$

we have $g(X; \sigma^2, n) = X/\sigma^2 C_n = Y$, and we see that $Y$ is distributed as $W(\chi^2; n)$ and is independent of any unknown parameters except $n$.

Also $f(n)$ is given by $\int_0^{f(n)} W(\chi^2; n)\, d\chi^2 = \beta$, and $h(\theta, n) = \sigma^2 C_n f(n)$.

Suppose in step one of our procedure we observe $u_1, u_2, \cdots, u_m$ which is a random sample of size $m$ from a normal population with variance $\sigma^2$. If we let

$$z = \sum_{i=1}^{m}(u_i - \bar{u})^2,$$

then since $z/\sigma^2$ is distributed as $W(\chi^2; m)$, it is clear that $P[(z/\sigma^2) > f_m] = \beta$, where $f_m$ is such that

$$\int_{f_m}^{\infty} W(\chi^2; m)\, d\chi^2 = \beta.$$

Hence $t(z) = z/f_m$, and since all the conditions in (2) are satisfied, the sample size for the desired length of the confidence interval is the smallest integral value of $n$ satisfying

$$\frac{f(n) \cdot C_n \cdot z}{f_m} \leqq d.$$

*Example* 2. Next, suppose it is desired to determine the sample size such ιhat an $\alpha$ confidence interval on the mean of a normal population will have width less than $d$ with probability $\beta^2$. Let $v_1, v_2, \cdots, v_n$ be a random sample of size $n$(to be determined) from a normal distribution with mean $\mu$ and variance $\sigma^2 = \theta^2$. If we let

$$s_n^2 = \frac{1}{n-1} \sum (v_i - \bar{v})^2,$$

then an $\alpha$ confidence interval on $\mu$ is

$$\bar{v} - \frac{t_0 s_n}{\sqrt{n}} \leqq \mu \leqq \bar{v} + \frac{t_0 s_n}{\sqrt{n}},$$

where $t_0$ is such that

$$\int_{t_0}^{\infty} U(t, n)\, dt = \frac{1-\alpha}{2},$$

where $U(t, n)$ is "Student's" distribution with $(n - 1)$ degrees of freedom. The length of the interval is

$$X = 2 \frac{t_0\, s_n}{\sqrt{n}}.$$

If we let

$$Y = g(X; \theta, n) = \frac{(n-1)s_n^2}{\sigma^2} = \frac{n(n-1)X^2}{4t_0^2 \sigma^2},$$

then $Y$ is distributed as a Chi-square variate with $(n - 1)$ degrees of freedom, and is independent of any unknown parameters except $n$.

If $W(\chi^2; n)$ is a Chi-square frequency function with $(n - 1)$ degrees of freedom, then $f(n)$ is given by

$$\int_0^{f(n)} W(\chi^2; n)\, d\chi^2 = \beta,$$

and

$$X = h(\theta, n) = 2t_0 \sigma \frac{\sqrt{f(n)}}{\sqrt{n(n-1)}}.$$

Suppose $u_1, u_2, \cdots, u_m$ is a random sample of size $m$ from a normal population with variance $\sigma^2$ which is available from step one of our procedure.

If we let

$$z = \sum_{i=1}^{m} (u_i - \bar{u})^2,$$

then we have $P[(z/\sigma^2) > f_m] = \beta$, where $f_m$ is such that $\int_{f_m}^{\infty} W(\chi^2; m)\, d\chi^2 = \beta$. Hence, $t(z) = (z/f_m)^{\frac{1}{2}}$, and since all the conditions in (2) are satisfied, the sample size for step two is the smallest integral value of $n$ satisfying

$$\frac{2t_0\sqrt{z}}{\sqrt{f_m}} \cdot \frac{\sqrt{f(n)}}{\sqrt{n(n-1)}} \leqq d.$$

It is interesting to compare the method in this paper with the method presented by Stein [1] for setting a confidence interval on the mean of a normal population with unknown variance.

The procedure presented by Stein is to select a two step sample. Suppose the sample in the first step is $u_1, u_2, \cdots, u_m$ and is taken from a normal population with mean $\mu$ and variance $\sigma^2$. An $\alpha$ confidence interval on $\mu$ is

$$\bar{u} - \frac{t_m\, s}{\sqrt{m}} \leqq \mu \leqq \bar{u} + \frac{t_m\, s}{\sqrt{m}},$$

where $s^2 = 1/(m-1) \sum (u_i - \bar{u})^2$ and $t_m$ is the appropriate value from "Students" distribution with $m - 1$ degrees of freedom. The width of the interval is $2t_m s/m^{\frac{1}{2}}$ and if this is less than the desired width $d$, no second step is required. If $2t_m s/m^{\frac{1}{2}} > d$, then $n$ additional observations $w_1, w_2, \cdots, w_n$ are taken where $n \geqq (4t_m^2 s^2/d^2) - m$, and the $\alpha$ confidence interval is

$$\bar{z} - \frac{t_m\, s}{\sqrt{m+n}} \leqq \mu \leqq \bar{z} + \frac{t_m\, s}{\sqrt{m+n}},$$

where

$$\bar{z} = \frac{n\bar{w} + m\bar{u}}{m+n}.$$

The width of the interval is $2t_m s/(m+n)^{\frac{1}{2}}$, and this is less than $d$.

It is to be noted that observations in the second sample are used *only* to compute the mean, $\bar{z}$.

Let us assume that the observations in the first step are taken from a normal population with mean $\mu_1$ and variance $\sigma_1^2$, and in the second step the mean is $\mu_2$ and the variance $\sigma_2^2$. Stein's method is valid if $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$. However, if $\mu_1 \neq \mu_2$, but $\sigma_1^2 = \sigma_2^2$, the method can still be used to set a specified confidence interval on $\mu_2$; the only alteration is that the second step requires a sample of size $n + m$ and $\bar{z}$ is the mean of this sample. That is to say, the sample mean from step one is not used in computing the interval. In this case if the inequality

$$\frac{4t_m^2\, s^2}{d^2} \leqq n$$

in Stein's procedure is compared with the inequality

$$\frac{2t_0\sqrt{z \cdot f(n)}}{\sqrt{f_m \cdot n(n-1)}} \leqq d$$

for the method presented in this paper, it is evident that Stein's procedure is to be preferred.

Next suppose that $\mu_1 \neq \mu_2$ and $\sigma_1^2 \neq \sigma_2^2$. Then Stein's procedure gives a confidence interval on $\mu_2$ with *known* probability (equal to 1) of a specified *width* but the confidence coefficient is *not known*. The method presented in this paper will give a confidence interval on $\mu_2$ with unknown probability of a specified width, but with known confidence coefficient.

Therefore, there may be cases when an experimenter would prefer the method in this paper over the one given by Stein for the mean of a normal distribution.

## REFERENCE

[1] CHARLES STEIN, "A two sample test for a linear hypothesis whose power is independent of the variance," *Ann. Math. Stat.*, Vol. 16 (1945), pp. 243–258.