Minimum-Width Confidence Intervals for Skewed Probability Densities

T. J. Fairclough

Institut Teknologi Brunei

Introduction

An element of choice arises when setting confidence limits with asymmetric probability densities. The requisite area, a, usually a = 0.05 or 0.01, can in principle be distributed over the two tails of the density in an infinity of ways. A common choice is to place half of the given area in each tail and then proceed to calculate the appropriate confidence interval on this basis with the aid of statistical tables or computer software. One drawback of this method is that it doesn't usually provide the smallest confidence interval, indeed intervals of slightly smaller width can sometimes be obtained by distributing the tail areas in different proportions by trial and error [Reference 1]. This situation does not arise with symmetrical distributions, where placing half of the area in each tail automatically achieves the confidence interval of minimum width.

In the first part of this article a simple procedure is described to obtain the minimum-width confidence interval using a skewed density passing through the origin. It is shown that the minimum width occurs when the two tails have equal heights. The procedure is then applied to a c^2 statistic for the common cases a = 0.05 and a = 0.01.

The c^2 distribution is also widely used in the calculation for setting confidence limits on the unknown population variance and standard deviation using a sample standard deviation statistic. It is generally desirable that the widths of these confidence intervals also be as small as possible. The problem of minimising these widths is considered in the latter part of this article.

Condition for Minimum Width

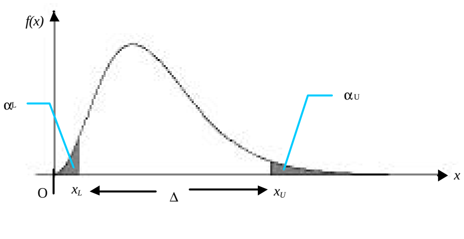

Figure 1

Figure 1 shows a skewed density function for the positive random variable X over an infinite range. As usual the probability that X takes on a value less than or equal to u is given by

$$P(X \le u) = \int_0^u f(x)dx$$

In Figure 1 the given constant area α has been arbitrarily divided into two parts; one part a_L has been inserted in the left-hand tail and the remaining part $a_U = a - a_L$ placed in the right-hand tail. The quantity $\Delta = x_U - x_L$ is therefore the width of the $(1-a)$*100% confidence interval for this particular choice of the partitioning of the given area a and it is required to find the smallest value of Δ amongst all such possible partitions of a.

The independent variable x_L is allowed to vary between zero (when all the area is in the right-hand tail) and some value x_{Lmax} (when all the area is in the left-hand tail). It is instructive to consider qualitatively the anticipated behaviour of Δ as x_L varies over this range. It is clear from Figure 1 that when x_L is zero Δ has some finite value which may be found by solving $1-a = \int_0^{\Delta} f(x)dx$. It is also clear that $\Delta \to \infty$ as $x_L \to x_{Lmax}$. It is possible that as Δ varies continuously between these two extremes there could be a value of Δ smaller than the starting value, if so there will be a minimum turning point on the graph of Δ v x_L

The requirement that the tail areas sum to α is expressed by

$$a = \int_0^{x_L} f(x)dx + \int_{x_L+\Delta}^{\infty} f(x)dx \quad [1]$$

Equation [1] defines Δ implicitly as a function of the independent variable x_L for any given constant value of a and with $0 \le x_L \le x_{Lmax}$. In the usual way Δ will have a minimum value if

$$\frac{d\Delta}{dx_L} = 0 \quad \text{and} \quad \frac{d^2\Delta}{dx_L^{\,2}} > 0 \quad [2]$$

Define the function

$$F(x) = \int_0^x f(u)du \quad [3]$$

Clearly $F(0) = 0$, $\frac{dF(x)}{dx} = f(x)$ and, since $f(x)$ is a density function, it follows that

$$\underset{N\to\infty}{Lim}[F(N)] = 1 \quad [4]$$

Using [1], [3] and [4] the condition that the two tail areas sum to a may now be expressed by

$$1 - a = F(x_L + \Delta) - F(x_L) \quad [5]$$

Differentiating both sides of [5] with respect to x_L gives

$$0 = f(x_L + \Delta)(1 + \frac{d\Delta}{dx_L}) - f(x_L) \quad [6]$$

so that

$$\frac{d\Delta}{dx_L} = \frac{f(x_L) - f(x_L + \Delta)}{f(x_L + \Delta)} \quad [7]$$

We observe that $\frac{d\Delta}{dx_L} = 0$ when

$$f(x_u) = f(x_L) \quad [8]$$

By differentiating [7] and using [8] we find that at the turning point

$$\frac{d^2\Delta}{dx_L^{\,2}} = \frac{g(x_L) - g(x_L + \Delta)}{f(x_L + \Delta)} \quad [9]$$

where $g(x) = \frac{df(x)}{dx}$.

For the type of situation illustrated in Figure 1 we see from [9] that the turning point will be a minimum since $g(x_L)$ is positive and $g(x_L + D)$ is negative.

The condition [8] is a necessary condition for the width to have a minimum value; it is automatically satisfied for symmetrical distributions when equal areas are placed in each tail. Satisfying the condition [8] simultaneously with the area requirement expressed by equation [1] enables the determination of the minimum width for a skew distribution with the given value of α. This is carried out in the next section using the c^2 density.

Minimum Width Confidence Intervals for c^2

If X is a chi-squared distributed random variable the appropriate density function is given by the χ^2 density with ν degrees of freedom [Reference 1]

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\,\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} & x > 0 \\ 0 & x \le 0 \end{cases} \quad [10]$$

The densities for n = 1 and ν =2 decrease monotonically so that $f(x_u) < f(x_L)$ and condition [8] cannot be satisfied. It can be shown that the smallest values of Δ for these cases are simply obtained by placing all of the given area α in the right hand tail.

For values of n >2 the condition that $f(x_u) = f(x_L)$ when applied to [10] leads to the condition that for the c^2 density

$$x_U - x_L = (n-2)\ln(\frac{x_U}{x_L}) \quad n > 2 \quad [11]$$

The required values of x_L and x_U associated with the stationary-width confidence interval for the given a may be found by using [11] together with the requirement that the areas in the two tails must sum to the given value of a as expressed by equation [12]

$$a = \int_0^{x_L} f(x)dx + \int_{x_U}^{\infty} f(x)dx \quad [12]$$

The simultaneous equations [11] and [12] may be solved numerically. A convenient procedure is to numerically solve equation [12], treated as a function of the single variable x_L, repeatedly using [11] where necessary to obtain and substitute the appropriate value of x_U corresponding to the current value of x_L. When the appropriate value of x_L has finally been found the corresponding value of x_U may be obtained using [11] and the stationary-width confidence interval for that value of a is then obtained from $\Delta = x_U - x_L$

Figure 2 shows the result of applying this procedure to the c^2 density with n = 5 degrees of freedom and a =0.05. For comparison the widths of all the possible 'trial' confidence intervals consistent with the given value of α are also shown as well as the minimum width of the confidence interval, Δ, calculated as described above. This provides a reasonably convincing demonstration that the simultaneous solution of Equations [11] and [12] does indeed give the minimum width without the need for trial and error.

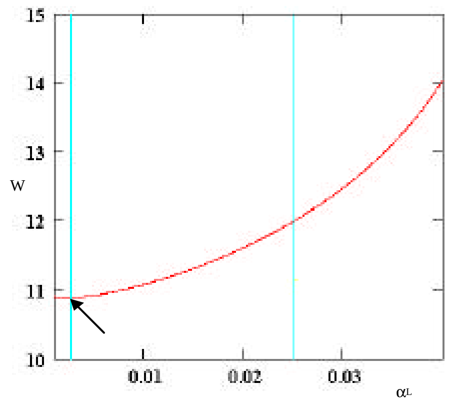

Figure 2

Figure 2 is a graph showing the width, W, of the 95% confidence intervals for a = 0.05 and n = 5 as a function of the area, a_L, placed in the left hand tail. The arrow marks the situation for the minimum width and is obtained by placing most of the area in the right hand tail. The right vertical line drawn at a_L = 0.025 shows the situation when equal areas are placed in each tail. At the minimum width a_L = 0.0023, x_L = 0.296 and x_U =11.191

Denoting the minimum width by Δ_{min} and the corresponding 'equal tail areas' width by d the quantity $P = \frac{\Delta_{\min} - d}{d}.100\%$ is plotted as a function of n > 2 in Figure 3. This shows the modest reductions in widths that may be obtained.

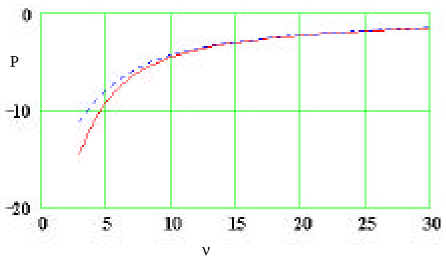

Figure 3

Figure 3 shows the percentage difference between the minimum width confidence interval and the corresponding width obtained when the areas in each tail are equal, plotted as a function of the number of degrees of freedom. The solid line is for a = 0.05 and the dotted line for a = 0.01.

Minimum-width confidence limits for the population variance and standard deviation

The standard procedure to estimate the confidence limits on the population variance, σ^2, and the standard deviation, σ, relies upon the fact that the quantity $X = \frac{(n-1)S^2}{s^2}$ is chi-squared distributed with ν = n-1 degrees of freedom. Here S is the given sample standard deviation calculated from the n experimental observations.

This implies that with $a_L + a_U = a$ we can be (1-a)*100% certain that

$$x_L \le \frac{(n-1)S^2}{s^2} \le x_U \quad [13]$$

In practice Equation [13] is often rearranged to provide the appropriate confidence limits on the variance or on the standard deviation and used either in the form of Equation [14] or [15] respectively:

$$\frac{(n-1)S^2}{x_U} \le s^2 \le \frac{(n-1)S^2}{x_L} \quad [14]$$

$$\frac{S\sqrt{n-1}}{\sqrt{x_U}} \le s \le \frac{S\sqrt{n-1}}{\sqrt{x_L}} \quad [15]$$

The required values of x_L and x_U are found from tables or software, usually by placing half of the area, a, into each tail of the c^2 density.

The width of the confidence interval for [13] is given by $\Delta = x_U - x_L$ and may thus be investigated by the method already described. The problem of finding the minimum widths of the confidence intervals for Equations [14] and [15] is considered below using Lagrange undetermined multipliers. Although this method only identifies the turning points it is found that in the cases of interest these do correspond to minimum values.

From [14] and [15] the widths of the confidence intervals for the variance and the standard deviation may be denoted respectively by

$$\Delta_{s^2} = S^2(n-1)\left(\frac{1}{x_L} - \frac{1}{x_U}\right) \quad [16]$$

and

$$\Delta_s = S\sqrt{(n-1)}\left(\frac{1}{\sqrt{x_L}} - \frac{1}{\sqrt{x_U}}\right) \quad [17]$$

[16] and [17] serve as new objective functions to be separately minimised subject to the constraint [5] that x_L and x_U must always define two tail areas with a fixed sum equal to the given α. It is important to note that the particular choice of x_L and x_U required to minimise Δ will not generally serve to also minimise either Δ_{s^2} or Δ_s. Indeed for the cases where ***n*** =1 and ***n*** = 2, the smallest value of Δ is achieved by placing all the area in the right hand tail thus setting x_L to zero. This choice leads to infinite widths for both Δ_{s^2} or Δ_s when the reciprocal of x_L is formed in [16] and [17] and is clearly an unsuitable choice. In this section we shall also implicitly include consideration of the cases where ***n*** =1 and ***n*** = 2

Consider first the problem of minimising Δ_{s^2}. With l as an undetermined multiplier we define the function

$$V = S^2(n-1)\left(\frac{1}{x_L} - \frac{1}{x_U}\right) + l\left(a - 1 - F(x_L) + F(x_U)\right) \quad [18]$$

The second term on the right of Equation [18] represents the fixed-area constraint on x_L and x_U corresponding to Equation [5] and the first term is the constrained function corresponding to the variance width to be minimised. Differentiating *V* partially with respect to x_L and x_U, setting the results to zero and eliminating λ yields the result

$$\frac{f(x_L)}{f(x_U)} = \left(\frac{x_U}{x_L}\right)^2 \quad [19]$$

Using the definition of *f(x)* for the chi-squared density in [10] we can now obtain the necessary condition for minimising the width of the confidence interval containing the variance as

$$x_U - x_L = (n+2)\ln\left(\frac{x_U}{x_L}\right) \quad [20]$$

It is interesting to note the similarity between [20] and [11]. In this context it may be observed that [11] could also have been derived in the same way as [20] but by starting with

$$V = x_U - x_L + l\left(a - 1 - F(x_L) + F(x_U)\right)$$

instead of the *V* used in [18]

[20] plays the same role in the minimisation of the variance interval Δ_{s^2} as is played by [11] in the minimisation of Δ.

We can also use this same multiplier technique to minimise Δ_σ, the width of the confidence interval for σ, by repeating the steps from [18] to [20] but this time starting with a new definition of *V* incorporating the expression for Δ_σ as given in [17]

$$V = S\sqrt{(n-1)}\left(\frac{1}{\sqrt{x_L}} - \frac{1}{\sqrt{x_U}}\right) + l\left(a - 1 - F(x_L) + F(x_U)\right) \quad [21]$$

The results analogous to [19] and [20] are then found to be

$$\frac{f(x_L)}{f(x_U)} = \left(\frac{x_U}{x_L}\right)^{3/2} \quad [22]$$

and

$$x_U - x_L = (n+1)\ln\left(\frac{x_U}{x_L}\right) \quad [23]$$

We are now in a position to use [20] and [23] together with [12] to investigate the possible reductions in the widths of the confidence intervals for the variance and the standard deviation as compared with the practice of dividing the area α equally between the two tails. Some results are illustrated in Figures 4, 5 and 6.

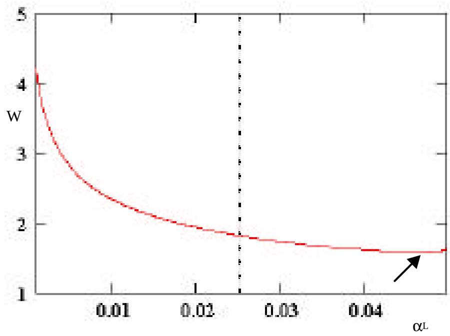

Figure 4

Figure 4 is a graph showing the width, W, of the 95% confidence intervals for the standard deviation with ***a*** = 0.05 and ***n*** = 5 as a function of the area, a_L, placed in the left hand tail. The sample standard deviation has been taken as unity. The arrow marks the situation for the minimum width found by simultaneously solving [23] and [12] and this time it is obtained by placing most of the area in the left-hand tail. The right vertical line drawn at a_L = 0.025 shows the situation when equal areas are placed in each tail. At the minimum width a_L = 0.0467, x_L = 1.109 and x_U = 17.743

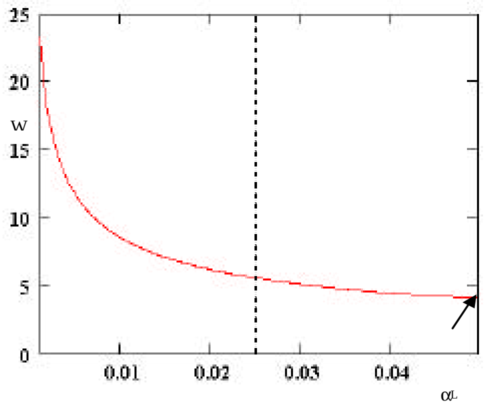

Figure 5

Figure 5 is a graph showing the width, W, of the 95% confidence interval for the variance with n = 5 as a function of the area, α_L, placed in the left hand tail. The sample variance has been taken as unity. Again the minimum width is obtained by placing most of the area in the left-hand tail and is found by simultaneously solving [20] and [12]. The right vertical line drawn at a_L = 0.025 shows the situation when equal areas are placed in each tail. At the minimum width a_L = 0.0494, x_L = 1.139 and x_U = 21.8.

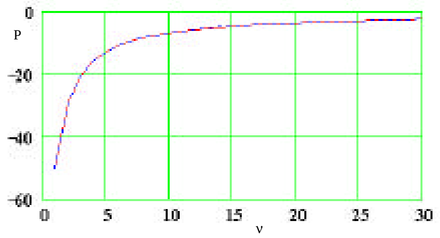

Figure 6

Figure 6 shows the percentage difference between the minimum width confidence interval for the standard deviation and the corresponding width obtained when the areas in each tail are equal, plotted as a function of the number of degrees of freedom for a = 0.05 and a = 0.01. With the vertical scale used there is very little difference between these two lines and in Figure 6 they are not resolved, giving the appearance of a single line.

Conclusion

Confidence intervals involving c^2 can be modestly reduced in width and in some extreme cases by as much as 50% over the ‘equal-area’ approach by using the methods described here. The scope for reduction decreases, as the number of degrees of freedom increases and the density becomes more symmetric. It is a matter of choice as to which confidence intervals are reported but it can sometimes be useful to be able to report narrower intervals for the population standard deviation if required particularly when only small sample sizes are available.

To avoid unnecessary calculations a set of reference tables may be drawn up and used to obtain the appropriate critical values of χ^2 in terms of ν and α that are required to obtain the minimum width confidence intervals for σ and σ^2 .

The techniques described here can also be applied to other suitable skew distributions.

References

‘Theory and Problems of Probability and Statistics’ by Murray. R. Spiegel. Schaum Outline Series, 1982.