Conference on Applied Statistics in Agriculture 1993 - 5th Annual Conference Proceedings

CAN SIMPLE RANDOM SAMPLING CONFIDENCE INTERVALS BE USED ON TRANSECT SAMPLING DATA?

William Noble

Can Simple Random Sampling Confidence Intervals be Used on Transect Sampling Data?

By William Noble

Kansas State University

Abstract

When sampling geographic regions, transect sampling may be easier and cheaper than simple random sampling. However, transect sampling data is more difficult to analyze. In the past, transect sampling data has sometimes been analyzed as if it was the result of simple random sampling. The purpose of this note is to present simulation results which show that this can lead to vastly inaccurate conclusions when one is calculating confidence intervals. In particular, an example is given of a purported 95% confidence interval which is actually a 49% confidence interval.

1 Introduction

When sampling geographic regions, transect sampling may be easier and cheaper than simple random sampling. However, transect sampling data is more difficult to analyze. In the past, transect sampling data has sometimes been analyzed as if it was the result of simple random sampling; see, for example, [2]. Young, Hammer, and Maatta in [3] discussed the assumptions which go into the calculations of confidence intervals for transect sampling data. In particular, they discussed assumptions of independence and normality and considered two techniques for calculating confidence intervals for means. Both techniques treat the data as if it were the result of simple random sampling. However, the first technique uses all of the data points, while the second only uses the transect means. The purpose of this note is to further pursue the issues raised in [3] and to show that in fact, the first technique can lead to conclusions with very low reliabillity. In particular, an example is given of a purported 95% confidence interval which is actually a 49% confidence interval.

2 A Transect Sampling Procedure

The simulations were based on sampling from a 1000×1000 grid ($1,000,000$ total possible sampling points). Fifteen transects are to be sampled, each transect having ten points with points spaced twenty units apart. The transect sampling procedure used in the simulations was the following.

1. An initial point (x_1, x_2) was selected at random from the 1000×1000 grid.
2. An angle θ was selected at random between 0° and 360°.
3. Ten sampling points were selected at 20 unit intervals from the grid, starting from (x_1, x_2) and going in the direction θ.
4. Sampling points were not allowed to be outside the grid. Thus, some transects may have had fewer than 10 points.
5. The above four steps were repeated 15 times to generate the 15 transects.

Using the terminology of [1], the *sampled population* consists of the $1,000,000$ points in the 1000×1000 grid. In practice, the sampled population is usually different from the *target population*. For example, for a 1000×1000 grid superimposed on a $1000m \times 1000m$ square field, the target population might be all possible points within the field, but the sampled population would only be the points in the grid. Any inferences which would be made in this case would be to the sampled population (the grid points), and not the target population.

3 Formal Calculation of 95% Simple Random Sampling Confidence Intervals

At each sampling point, a quantity is measured. For the purposes of the simulations, the quantity measured was

$$\begin{aligned} Y(x_1, x_2) &= 7, \ x_1 \leq 700 \\ &= 8, \ x_1 > 700. \end{aligned}$$

The quantity $Y(x_1, x_2)$ might, for example, correspond to the surface pH at the point (x_1, x_2). With this interpretation, the surface pH has the value 7 at 70% of the grid points, and has the value 8 at the other 30% of the grid points.

Suppose that there are n total sampling points, where n is a number less than or equal to 150. Each of the n sampling points generates a corresponding $Y(x_1, x_2)$. Labeling these values as $Y_1, Y_2, \ldots, Y_n$, it is formally possible to calculate a 95% simple random sampling confidence interval by using the usual formulas:

$$\overline{Y} \pm \frac{1.96s}{\sqrt{n}}.$$

The question to be answered in the next section is about the validity of this confidence interval, i.e., is it really a 95% confidence interval?

4 Using Simulation to Test the Validity of the Simple Random Sampling Confidence Interval

A valid 95% confidence interval should contain the true mean at least 95% of the time. The true mean for the "surface pH" in Section 3 is

$$\mu = (0.7 \times 7) + (0.3 \times 8) = 7.3.$$

To evaluate the validity of the 95% simple random sampling confidence interval calculated in the Section 3, the following simulation procedure was employed.

1. The transect sampling procedure described in Section 2 was used to generate a collection of $n \leq 150$ sampling points.
2. A formal 95% confidence interval was calculated.
3. It was determined whether or not this formal confidence interval contained the true mean $\mu = 7.3$.
4. The procedure described in the first three steps was repeated 40,000 times.
5. The proportion of the 40,000 times that the formal confidence interval contained the true mean was calculated.

5 Simulation Results and Conclusions

The results of the simulations were that the 95% simple random sampling confidence intervals contained the true mean in 49% of the simulations. In other words, a confidence interval which is supposed to contain the true mean 95% of the time actually only contained the true mean about half of the time.

The obvious lesson to be drawn from this example is that it is extremely hazardous to treat transect sampling data as if it were simple random sampling data when calculating confidence intervals. The simulation results obtained above lead to two questions.

First, although simple random sampling confidence intervals performed poorly for the function $Y(x_1, x_2)$ used in the simulation, there are other functions for which simple random sampling confidence intervals will do better. S. Sly at Kansas State University is currently investigating the performance of simple random sampling confidence intervals for various functions $Y(x_1, x_2)$.

Second, is it possible to obtain at least approximately valid confidence intervals from transect sampling data when one does not know the form of the function $Y(x_1, x_2)$? One obvious improvement is to use transect means instead of the individual sampling points. This, of course, involves some information loss, but does regain the validity of the confidence level. Additional simulations were performed by S. Sly which used transect means and the confidence interval formula $\overline{Y} \pm \frac{1.96s}{\sqrt{15}}$, where s is the sample standard deviation of the 15 transect means. These simulations resulted in confidence intervals which contained the mean 91% of the time. In general, such confidence intervals will be valid to the extent that the Central Limit Theorem applies.

Another possible approach is to view transect sampling as a special case of cluster sampling. In this case, one could perhaps first estimate "intracluster correlation" and then modify known formulas for estimator variance found in [1]. This approach is currently under investigation by the author.

References

[1] William G. Cochran. *Sampling Techniques.* John Wiley & Sons, New York, 3^{rd} edition, 1977.

[2] J.C. Powell and M.E. Springer. Composition and precision of classification of several mapping units of the Appling, Cecil, and Lloyd series in Walton County, Georgia. *Soil Science Society of America Proceedings*, 29:454–458, 1965.

[3] Fred J. Young, R. David Hammer, and Jon Maatta. Confidence intervals for soil properties based on differing statistical assumptions. In *Proceedings of the 1992 Kansas State University Conference on Applied Statistics in Agriculture*, pages 87–103, Manhattan, KS, 1992.

Department of Statistics
Kansas State University
Manhattan, KS 66506-0802