A JACKKNIFE-DERIVED VISUAL APPROACH FOR SAMPLE SIZE DETERMINATION

UN APPROCCIO GRAFICO DERIVATO DAL JACKKNIFE PER LA DETERMINAZIONE DELLE DIMENSIONI DEL CAMPIONE

Roberto Confalonieri

Department of Crop Science, Section of Agronomy, University of Milano, Via Celoria 2, 20133 Milano, Italy.

Abstract
The determination of the sample size is one of the most important topics in field research and preliminary sampling sessions should be carried out to determine the size of the sample before the collection of real data. A new method for sample size determination is proposed, based on a jackknife visual evolution. The method has been tested to determine (i) the sample size in a rice field in order to collect aboveground biomass data and (ii) the minimum number of time domain reflectometry (TDR) data acquisitions necessary to get a reliable estimation of soil water content in a maize field. The 18-plants sample size determined for rice aboveground biomass with the proposed method is coherent with the 20-plants samples traditionally harvested for field experiments with rice; while the 8 acquisitions indicated for soil water content take into account both the effort required for collecting data and the measurement accuracy.
The method does not care if the sample units are normally distributed or not and takes into account both the variation of accuracy for increasing sample sizes and, indirectly, the effort required for getting the measurements.

Keywords: Monte Carlo, rice*, Oryza sativa L.*, resampling, statistical inference, number of plants, aboveground biomass, soil water content, TDR

Riassunto
La determinazione delle dimensioni del campione è uno degli aspetti più importanti nella ricerca di campo: sessioni di campionamento preliminari mirate alla sua valutazione dovrebbero sempre essere condotte prima della raccolta dei dati veri e propri. In questo lavoro viene proposto un nuovo metodo per la determinazione delle dimensioni del campione, basato su un evoluzione grafica del jackknife. Il metodo è stato testato (i) per la raccolta di dati di biomassa per la coltura del riso e (ii) per la determinazione del numero minimo di letture al TDR (time domain reflectometer) necessarie per una stima affidabile del contenuto idrico del terreno in un campo di mais. La dimensione del campione di 18 piante indicata dal metodo proposto è coerente con il campione di 20 piante tradizionalmente raccomandato per esperimenti di campo in risaia. Le 8 acquisizioni al TDR indicate per il contenuto idrico del terreno tengono in considerazione sia lo sforzo necessario per la raccolta del dato che l'affidabilità della misura.
Il metodo funziona sia per campioni distribuiti normalmente che non e, indirettamente, tiene conto dello sforzo necessario per ottenere il dato sperimentale.

Parole chiave: *Monte Carlo, riso, Oryza sativa L., ricampionamento, inferenza statistica, numero di piante, biomassa, contenuto idrico del terreno, TDR*

Introduction
The determination of the sample size is one of the most important topics for the collection of reliable measurements of variables describing plant growth (e.g. aboveground biomass (AGB), leaf area index) and, in general, for all the aspects of field research.
In many cases, the classical inferential methods based on the t distribution for determining the sample size cannot be used: the characteristics of the population (μ and σ) are unknown. They varies according to many factors (e.g. homogeneity of the incorporation of residues, of sowing, of germination, of emergence, etc.). For these reasons, it is impossible to have a presumptive knowledge of the sampling error that can be accepted for the calculation of the sample size: it could change from an experimental field (or situation) to another. For this reason, it is better to analyze the trend of the relative variability inside the same field when the sample size increases rather than assuming absolute criterions for the acceptance of a specific variability for the determination of the sample size. In this way, the effort required for processing a sample can be easily taken into account too. Moreover, a reliable evaluation of the fact that the population is normally distributed is often not possible because, above all for time-consuming or costly methods, only

Corresponding author. tel. +39 02 5031 6592
fax +39 02 5031 6575
e-mail address: roberto.confalonieri@unimi.it (R.Confalonieri).

few observations can be determined in the preliminary study carried out for determining the sample size.
It is possible to find in the Literature examples of alternative methods for sample size determination. Wolkowski *et al.* (1988), in an experiment aiming at comparing field plot techniques for corn grain and dry matter yield estimation, determined the sample size by harvesting all the plants belonging to a row and determining AGB on 1, 5, 10, 15, 20, 25, all plants of the row randomly extracted from the whole harvested plants and calculating the coefficient of variation [CV = standard deviation (average value)$^{-1}$] of each sample. They found that the coefficient of variation minimizes for 10 and 15 – plants samples. The samples extracted from the same row can not be considered completely independent. Moreover, the Authors don't specify if the plants belonging to a sample had been used to determine the AGB of other samples.
It's possible to find in the Literature studies for rice sample size (= number of plants) determination, although they were not directly related to AGB determination. For example, Tirol Padre *et al.* (1988), in an experiment about acetylene reduction activity in different rice varieties, noticed that the average CV (computed from the different varieties CV, obtained from many measurement replicates) decreased from 33% with 1 plants to 11% for 6 plants. The Authors discussed that, for some samples, the variances were correlated to the means (not-homoscedastic) and for others not (homoscedastic). For these reason, probably the CV is not the better parameter for sample size determination: it should be more correct the separated analysis of means and standard deviations.
A different approach, based on the different components of the effort required for the whole process to determine an experimental value, was proposed by Yonezawa (1985) to determine the minimum number of plants required for the conservation of plant genetic resources. The Author concluded that (i) a sample size as small as 10 plants per site or population is reasonable to cover a large target area and (ii) it is more important to analyze a wide number of site or population. This conclusion could be extendable to an experimental field by considering, for example, the sites corresponding to the plots per treatment. In this way, it's possible to conclude that, for the same available effort, it's better to harvest few plants per plot but dispose of more plots for each treatment.
From the 60s resampling methods are increasingly diffused in biological and environmental sciences. They conceptually derive from the "Monte Carlo" methods, for the first time applied to a physical process by Barker (1965) but they are based on the repeated use of the data from the same sample (the only one sample which has been collected). Efron and Tibshirani (1991) underline the enormous potentiality of these techniques, which widely use the calculation capability of computers, for the research on inferential statistics. The bootstrap (Efron, 1979; Efron and Tibshirani, 1993) and the jackknife (Quenouille, 1949; Tukey, 1958) are particularly important, both because they are increasingly used and because of the number of studies and developments they generated.
Therefore, the objectives of this study are (i) the development of an alternative method for sample size determination based on an evolution of the jackknife; (ii) its evaluation for determining the sample size for aboveground biomass in a rice field and for soil water content in a maize field. In the first case the sample size will be determined as number of plants and, in the second, as number of TDR (Time Domain Reflectometer) measurements.

Materials and methods

Experimental data

Experimental data were collected in 2 experiments. The first was carried out in Besate (northern Italy, latitude 45° 18' N, longitude 8° 58' E) during 2003. Rice (*Oryza sativa* L; cv. Volano) was sown in April 28 and grown under flooded conditions. No water stresses were observed during the crop cycle and the management has allowed to prevent the presence of weed and pests. During this experiment, the studied variable was aboveground biomass. The samples were collected when the plants were young (3 leaves) and the dry weight of each plant was very low. For this reason, we have chosen to use a group of 3 plants as sampling unit instead of a single plant in order to avoid errors due to the measurement of very small quantities of biomass.
The second experiment was carried out in Lodi Vecchio (northern Italy, latitude 45° 19' N, longitude 9° 26' E) during 2004. Maize (*Zea mais* L) was grown under optimal water and nutrients availability. The soil is a Ultic Haplustalfs fine silty, mixed, mesic (Soil Survey Staff, 1999), subacid, with medium organic matter content, sufficient available phosphorous and medium potassium content. The studied variable was soil water content. In this case, the sampling unit was a single TDR (Time Domain Reflectometer) measurement event.

A visual jackknife evolution for determining the sample size

The proposed method can be considered a modification of the jackknife, developed by Tukey (1958) at the end of the 50s on the basis of an idea of Quenouille (1949; 1956) of some years before and reviewed, with its recent developments, by Hinkley (1983).

According to this technique, the original sample of N elements is divided into groups of k elements. If N is low, k may be equal to 1. $\frac{N!}{(N-k)!k!}$ virtual samples (combinations without repetitions) of (N-k) elements are generated by eliminating $\frac{N!}{(N-k)!k!}$ times k different values from the original sample. Mean, standard deviation, etc. can be calculated for all the generated virtual samples.
In the presented method, different values of k are used. The process of generation of the $\frac{N!}{(N-k)!k!}$ virtual samples is repeated (N-1) times with k assuming values from

1 to (N-2), for a total of $\sum_{k=1}^{N-2} \frac{N!}{(N-k)!k!}$ different virtual samples.

Mean and standard deviation are computed for all the generated samples and plotted on two charts, with the values of (N-k) on the X-axis (Figures 1 and 2) and the means (or standard deviations) on the Y-axis, in order to get a visual representation of how the means and the standard deviations of the generated samples vary with increasing sample size.

Analyzing the trends of the means and the standard deviations for increasing values of (N-k), the sample size is considered equal to the (N-k) value for which the variability among the means and the standard deviations computed for almost all the generated samples changes its decreasing rate, assuming a slow – asymptotic trend for increasing (N-k) values. This change of rate can be determined by the observation of how the variability among means and standard deviations evolves for increasing (N-k) values or by using an automatic algorithm. In the second case, the following procedure is adopted:

1. select a (N-k) value (N-k)' higher than 2 and lower than N-2;
2. for i from 2 to (N-k)', calculate mean and standard deviation of the 5% highest means (respectively Y_i and s_i);
3. compute a linear regression weighted on the standard deviations assuming as independent variable the (N-k) values and as dependent one the Y_i values and computing the R^2 of the regression (R^2_1);
4. repeat the steps [2] and [3] for the 5% lowest means, determining R^2_2;
5. for i from (N-k)' to N-1, repeat the steps from [2] to [4] to compute R^2_3 and R^2_4, imposing that the slope of the regression is 0;
6. calculate SR^2 by adding R^2_1, R^2_2, R^2_3 and R^2_4;
7. repeat the steps from [2] to [6] for each of the (N-k)' values higher than 2 and lower than N-2;
8. the automatically determined sample size based on the means (AM) is the (N-k)' value corresponding to the lower SR^2.

The same procedure is used to determine the sample size basing on the standard deviations (ASD). The highest between AM and ASD is the automatic sample size.

Two important features of the presented method for the sample size determination are: (i) the initial sample may be constituted both by independent and not-independent data and (ii) the data may be both normally or not-normally distributed. Moreover, it supplies a visual, easy-to-analyze representation of a particular sample size directly related to its variability and to the variability of bigger and smaller samples sizes. In this way, it is easier to consider the problem of sample size determination both from the statistical point of view and from the point of view of the available effort for carrying out the experimentation.

Application of the proposed method for determining the sample size

For the first experiment, AGB of 12 3-plants samples (36 plants; aggregated sample of 3 sub-samples) was measured.

250 dispositions without repetitions were generated through a randomization process using the 12 samples. Only 250 dispositions were randomized because, when the initial sample is not numerous, an increase of the re-samplings number does not correspond to an effective improvement of the estimation (Manly, 1991). The obtained values were collected in a matrix (Matrix O; 12 rows; 250 columns) with the dispositions in the columns and the values of each disposition in the rows. A new matrix was created (Matrix M; 11 rows; 250 columns) and, for each of its columns, the mean value computed on the first 2 elements of correspondent column of the Matrix O was stored in the first row; the mean value computed on the first 3 elements in the second rows and so on. The same procedure used to create the Matrix M was followed to create a Matrix SD, containing the standard deviations of the values stored in the Matrix O. Therefore, the elements of Matrix M and Matrix SD are computed by using the following equations.

$$MatrixM_{nj} = \frac{\sum_{i=1}^{n} MatrixO_{ij}}{n}$$

$$MatrixSD_{nj} = \sqrt{\frac{\sum_{i=1}^{n} \left(MatrixO_{ij} - \frac{\sum_{i=1}^{n} MatrixO_{ij}}{n} \right)^2}{n}}$$

where j indicates the dispositions and n is equal to (N-k).

The elements of Matrix M and Matrix SD were plotted separately with the 250 series corresponding to the columns of the two matrixes (Figure 1 and 2).

The same procedure was followed for determining the sample size (number of TDR measurement events) for soil water content determination (second experiment). The only two differences are that the sampling unit was a single TDR data acquisition instead of an aggregated sample of 3 plants and that N was, in this case, 18 instead of 12 (Figure 3 and 4).

Jack: a software for sample size determination using the visual jackknife

A software was created for the application of the proposed method (Figure 5). It generates the virtual samples and the matrixes of the means and of the standard deviations. It produces the relative diagrams, draws the 4 regression lines used by the automatic method for sample size determination and compute the CV obtainable with the automatic method. It gives to the user the possibility of choosing different sample sizes allowing him to compare the CVs obtainable with the two methods (automatic or not). The software is free downloadable at the web site:

http://users.unimi.it/agroecol/confalonieri.php.

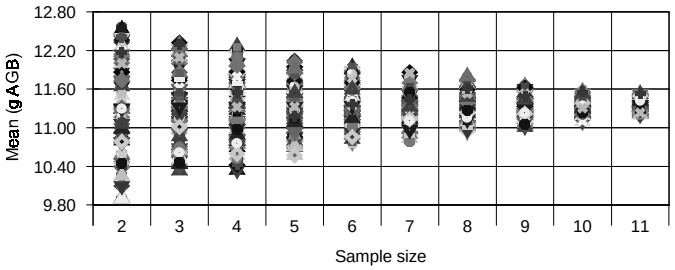

Fig. 1 – Aboveground biomass (rice) - Means of the generated populations when the jackknife is applied for different k values. (N-k) values on the X-axis, with k from (N-2) to 1. More details in the text. k is the number of observations not used by the jackknife; N is the total number of observations

Fig. 1 - Biomassa aerea (riso) - Medie delle popolazioni generate applicando il jackknife per differenti valori di k. Sull'asse delle X i valori (N-k), con k da (N-2) a 1. Maggiori dettagli nel testo. k è il numero di osservazioni non utilizzate nel jackknife; N è il numero totale di osservazioni

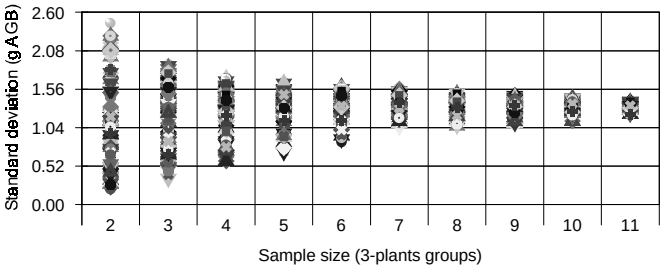

Fig. 2 – Aboveground biomass (rice) - Standard deviations of the generated populations when the jackknife is applied for different k values. (N-k) values on the X-axis, with k from (N-2) to 1. More details in the text. k is the number of observations not used by the jackknife; N is the total number of observations

Fig. 2 - Biomassa aerea (riso) - Deviazioni standard delle popolazioni generate applicando il jackknife per differenti valori di k. Sull'asse delle X i valori (N-k), con k da (N-2) a 1. Maggiori dettagli nel testo. k è il numero di osservazioni non utilizzate nel jackknife; N è il numero totale di osservazioni

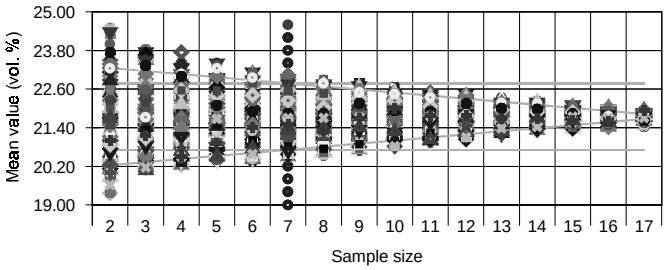

Fig. 3 – Soil water content - Means of the generated populations when the jackknife is applied for different k values. (N-k) values on the X-axis, with k from (N-2) to 1. More details in the text. k is the number of observations not used by the jackknife; N is the total number of observations.

Fig. 3 - Contenuto idrico del terreno - Medie delle popolazioni generate applicando il jackknife per differenti valori di k. Sull'asse delle X i valori (N-k), con k da (N-2) a 1. Maggiori dettagli nel testo. k è il numero di osservazioni non utilizzate nel jackknife; N è il numero totale di osservazioni

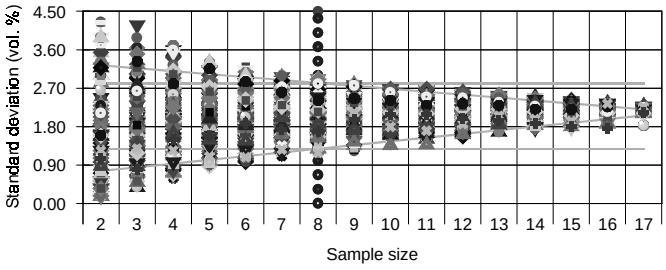

Fig. 4 – Soil water content - Standard deviations of the generated populations when the jackknife is applied for different k values. (N-k) values on the X-axis, with k from (N-2) to 1. More details in the text. k is the number of observations not used by the jackknife; N is the total number of observations.

Fig. 4 - Contenuto idrico del terreno - Deviazioni standard delle popolazioni generate applicando il jackknife per differenti valori di k. Sull'asse delle X i valori (N-k), con k da (N-2) a 1. Maggiori dettagli nel testo. k è il numero di osservazioni non utilizzate nel jackknife; N è il numero totale di osservazioni

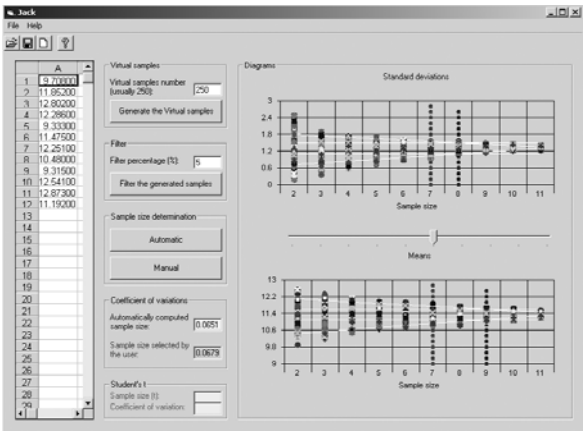

Fig. 5 – Jack: a software for the visual-jackknife application (http://users.unimi.it/agroecol/confalonieri.php). User's interface

Fig 5 - Jack: un software per l'applicazione del visual-jackknife (http://users.unimi.it/agroecol/confalonieri.php). Interfaccia utente

Results and discussion

Aboveground biomass

Figure 1 shows that the differences between the means of the generated populations, obtained with the proposed method, decrease for sample size equal to (6×3) plants. These differences continue to slightly decrease with higher sample sizes but 6 groups of three plants can be considered a satisfactory compromise between the effort to get the measure and its reliability. In fact, although the differences between the average values decrease also for sample sizes higher than the indicated one, there is a clear change in the height of the region of the plot engaged by the 250 series.

The same considerations can be discussed for Figure 2 (standard deviations of the generated populations). It's possible to notice that the height of the region of the plot engaged by the 250 series is almost constant for sample sizes higher than 6 group of 3 plants, while higher difference in standard deviations can be observed considering less than 18 plants. In practice, a higher number of plants would increase the effort required by the measurement without a comparable increase in the accuracy.

In this example, the sample size was determined by graphically analyzing the two diagrams. This "visual" procedure allows to determine sample sizes which take into account the effort required to get data and the experimenter's judgement.

Soil water content

For the determination of the number of TDR measurement events required to get a reliable soil water content estimation, the automatic procedure was used (see the section "Materials and methods").

Figure 3 and 4 show, respectively, the means and the standard deviations of the generated populations. The grey straight lines represent the linear regressions on which the R^2 are computed. The intersection between the oblique lines and the horizontal ones represents the automatically computed samples size, also indicated by the vertical series of dots. It is possible to notice that the automatically computed sample size is 7 in Figure 3 (means) and 8 in Figure 4 (standard deviations). In this kind of cases, the proposed algorithm advise to chose the highest.

The used algorithm has been planned to reproduce the behavior of an experimenter in selecting a sample size and it is particularly useful when the area of the diagrams engaged by the generated series decreases, for increasing sample sizes, too regularly to allow the experimenter to confidently notice discontinuities.

Conclusions

When parametric modelling and theoretical analysis are difficult, the bootstrap (Efron, 1979; Efron and Tibshirani, 1993) and the jackknife (Quenouille, 1949; Tukey, 1958) are good alternatives for analyzing the characteristics of a population (Park and Willemain, 1999).

The proposed method, based on a visual evolution of the jackknife has shown to be reliable for sample size determination, determining sample sizes of (i) 18 plants (6 aggregated samples of 3 plants) for rice aboveground biomass determination and (ii) 8 TDR measurement events for soil water content estimation in a maize field. The first value is coherent with the 20 plants per plot recommended to be harvested by Gomez (1972) in his manual about field experiments with rice and traditionally used in field experiments.

Future studies will evaluate the applicability of the proposed method to other variables (e.g. leaf area index, soil nitrogen content, etc).

References

Barker, A.A., 1965. Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. Australian Journal of Physics, 18, 119-133.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Annals of statistics, 7, 1-26.

Efron, B., Tibshirani, R., 1991. Statistical data analysis in the computer age. Science, 253, 390-395.

Efron, B., Tibshirani, R., 1993. An introduction to the bootstrap. Chapman & Hall, New York.

Gomez, K.A., 1972. Techniques for field experiments with rice: layout, sampling, sources of error. International Rice Research Institute, Los Baños, Philippines, pp. 46.

Hinkley, D.V., 1983. Jackknife methods. Encyclopedia of Statistical Science, 4, 280-287.

Manly, B.F.J., 1991. Randomization and Monte Carlo Methods in Biology. Chapman and Hall, London, New York, USA, pp. 281.

Park, D., Willemain, T.R., 1999. The threshold bootstrap and threshold jackknife. Computational Statistics & Data Analysis, 31, 187-202.

Quenouille, M.H., 1949. Approximate tests of correlation in time series. Journal of The Royal Statistical Society, Series B, 11, 68-84.

Quenouille, M.H., 1956. Notes on bias in estimation. Biometrika, 43, 353-360.

Soil Survey Staff, 1999. Soil Taxonomy, 2nd edition. Agric. Handbook n. 436, USDA-NRCS, 869 pp.

Tirol Padre, A., Ladha, J.K., Punzalan, G.C., Watanabe, I., 1988. A plant sampling procedure for acetylene reduction assay to detect rice varietal differences in ability to stimulate N2 fixation. Soil Biology and Biochemistry, 20, 175-183.

Tukey, J.W., 1958. Bias and confidence in not quite large samples (Abstract). Annals of Mathematical Statistics, 29, 614.

Wolkowski, R.P., Reisdorf, T.A., Bundy, L.G., 1988. Field plot technique comparison for estimating corn grain and dry matter yield. Agronomy Journal, 80, 278-280.

Yonezawa, K., 1985. A definition of the optimal allocation of effort in conservation of plants genetic resources with application to sample size determination for field collection. Euphytica, 34, 345-354.