DETERMINATION OF OPTIMUM SAMPLING FRACTION FOR
BOOTSTRAP RESAMPLING

DETERMINAÇÃO DE FRAÇÃO AMOSTRAL ÓTIMA PARA
REAMOSTRAGENS BOOTSTRAP

João Gabriel Malaguti¹
Samuel Faria Candido²

Abstract: Lately, there has been much discussion on the bootstrap resampling method, both as a way of estimating standard error and as a way of improving estimations with access to only one sample. However, little is found in literature discussing the size the bootstrap sample should take. This study aims to determine the existence of an optimum sampling fraction for resampling, analysing different estimators and number of resamples. An optimum fraction exists if, and only if, for every estimator and every amount of resamples, a fraction (or region) performs better in every population. Ten random populations were created by adding together different normal, Poisson and exponential distributions such that their means and variances are diverse. A Monte Carlo simulation with ten thousand iterations was done, taking random systematic samples from the populations and from these, bootstrap samples to estimate the mean, variance and respective standard errors. Results show the inexistence of a single optimum fraction. However, it does point to an optimum region for standard error estimation above 37.5%.

Keywords: Bootstrap resampling. Monte Carlo simulation. Systematic sampling.

Resumo: Ultimamente, muito tem-se discutido sobre o método bootstrap para reamostragens, tanto como método para a estimativa de erro padrão de estimadores quanto para melhorar a própria estimativa com apenas uma amostra. No entanto, pouco se encontra na literatura discutindo o tamanho que essa amostra bootstrap deve tomar. Este estudo tem como alvo determinar a existência de uma fração amostral ótima para reamostragem, analisando diferentes estimadores e quantidades de repetições. Uma fração ótima existe se, e somente se, para todos os estimadores e para todas as quantidades de repetições, uma fração (ou região) se mostra melhor em qualquer população. Foram criadas 10 populações aleatórias, somando diferentes distribuições normais, Poisson e exponenciais, de modo que suas médias e variâncias fossem diversas. Uma simulação Monte Carlo com dez mil iterações foi feita, retirando-se amostras sistemáticas da população e destas, amostras bootstrap para estimar-se a média, a variância e seus erros padrão. Resultados apontam a inexistência de fração ótima única, mas para a estimativa do erro padrão, temos uma faixa ótima acima de 37,5%.

Palavras-chave: Reamostragem bootstrap. Simulação Monte Carlo. Amostragem sistemática.

¹ Bacharel em Estatística. Universidade Federal de Juiz De Fora, Departamento de Estatística - Juiz de Fora, MG, Brasil. joaogmalaguti@gmail.com.
² Bacharel em Estatística. Universidade Federal de Juiz De Fora, Departamento de Estatística - Juiz de Fora, MG, Brasil. samuelfariacandido@gmail.com.
1 INTRODUCTION

Although literature on bootstrap resampling is relatively plentiful, little is found on the size of the resample. Even though it is important to non-parametric bootstrap, the size used in an ad hoc manner is the size of the original sample. In this study, we search for the optimum sampling fraction (through grid search) for different populations. By hypothesis, we have three scenarios of interest:

- The optimum sampling fraction is bigger than 100%, which means we need to enlarge the sampling fraction in order to better the estimations;
- The optimum sampling fraction is smaller than 100%, which means not only can we gain precision by restricting the sample, we also get smaller computing costs for the same number of replications;
- The optimum sample size is equal to 100%, therefore we have basis for using it.

Using Monte Carlo simulations (MC), we can determine the existence of such a single fraction or, in case it does not exist, secondary optimum fractions. These are optimum for some class though not all, for example the class of variance estimators or the class of standard errors estimators.

2 DEVELOPMENT

2.1 Studies

The main discussion on the size of bootstrap resample is found in Davison and Hinkley (1997), who point out that if the sampling fraction $f \ll 1$ (or 100%) the statistics of the resamples may be less stable than those based on resamples maintaining the original size.

Non-parametric bootstrap resampling, as defined by Efron & Tibishirani (1994), is simple to implement computationally, generating B simple random samples with reposition (SRSWR) from the original sample, such that these resamples are independent among each other. It is important to note that although simple, it is a very intensive process, even more so when part of a simulation study that involves many replications.
A Monte Carlo simulation study is, by definition, a method based on repeated samples made in computers in order to obtain approximate answers to probabilistic problems, being used since the beginning of the XX century (LEWIS & ORAV, 1989).

Using the statistical software R (R CORE TEAM, 2019), we created 10 populations by adding together normal, Poisson and exponential distributions with integer parameters chosen randomly between 2 and 20. We defined the original sample size $n = 1000$, four amounts of bootstrap resamples $B = \{500, 1000, 2000, 5000\}$, sixteen different sampling fractions $f = \{12.5, 25, 37.5, 50, 62.5, 75, 87.5, 100, 112.5, 125, 137.5, 150, 162.5, 175, 187.5, 200\}$ (in percentages), and number of Monte Carlo iterations $k = 10000$. As parameters, we chose the sample mean, the sample variance and their standard errors (SEs) for a sample of size $n$.

In every iteration, a systematic sample is taken from a population and from the sample $B$ resamples of size $n \cdot f$ are taken using SRSWR, which are then used in the estimation of the parameters chosen. At the end of all iterations for all possible combinations of $B, f$ and populations, we calculate the mean squared error (MSE) of the estimates. The MSE is defined as

$$MSE(y) = (y - Y)^2,$$

in which $y$ is the estimate for the parameter $Y$.

For the calculation of the MSE for the SEs, we use as parameter the estimates resulting from a MC simulation with a hundred thousand iterations, since there does not exist formal equations for the calculus of standard errors for a systematic sample. Finally, the MSEs were standardized by populations in order to remove their effects. The results are presented in the following graphs.

### 2.2 Results

The heat maps show the standardized MSE, by population (x-axis), sampling fraction (y-axis), parameter (mean or variance, respectively first and second columns) and class (estimator or standard error, respectively first and second rows). The amount of resamples is different for each graph (500 for Figure 1, 1000 for Figure 2, 2000 for Figure 3 and 5000 for Figure 4).
Figure 1 – Standardized mean square error by population, sampling fraction, parameter and class for 500 resamples.

Source: Elaborated by the authors.

Figure 2 – Standardized mean square error by population, sampling fraction, parameter and class for 1000 resamples.

Source: Elaborated by the authors.
Figure 3 – Standardized mean square error by population, sampling fraction, parameter and class for 2000 resamples.

Source: Elaborated by the authors.

Figure 4 – Standardized mean square error by population, sampling fraction, parameter and class for 5000 resamples.

Source: Elaborated by the authors.

Revista Mundi Engenharia, Tecnologia e Gestão. Paranaguá, PR, v.5, n.2, p. 215-01, 215-09, 2020
DOI: 10.21575/25254782rmetg2020vol5n21140
Observing the graphs, we can notice an absence of a pattern in the MSEs of the estimator both mean and variance that behave differently, not only among each other but also for different amounts of resamples.

Also clear is the existence of an optimum band for SE estimation, from 37.5% (sample size equal to 375) to, at least, 200%. We have then, a scenario of interest, since the optimum sampling fraction is smaller than 100%, meaning that not only we gain precision by restricting the sample; we also lessen our computing costs for a same number of resamples. This is seen both in the standardized values and in the squared standardized values (Figures 5, 6, 7 and 8).

When standardizing the values, that is, subtracting their mean and dividing by their standard deviation, the results are values centred around zero, both positive and negative values. In order to return the scale of the MSEs to a one direction scale and to better the visualization, these values were squared.

Figure 5 – Standardized mean square error squared by population, sampling fraction, parameter and class for 500 resamples.

Source: Elaborated by the authors.

Revista Mundi Engenharia, Tecnologia e Gestão. Paranaguá, PR, v.5, n.2, p. 215-01, 215-09, 2020
DOI: 10.21575/25254782rmetg2020vol5n21140
Figure 6 – Standardized mean square error squared by population, sampling fraction, parameter and class for 1000 resamples.

Source: Elaborated by the authors.

Figure 7 – Standardized mean square error squared by population, sampling fraction, parameter and class for 2000 resamples.

Source: Elaborated by the authors.
Figure 8 – Standardized mean square error squared by population, sampling fraction, parameter and class for 5000 resamples.

Source: Elaborated by the authors.

It is interesting to note that even in a one direction scale, there are no clear patterns for the estimators be it between them for a same number of resamples or even among different amount of resamples. Even their worst values (whites, yellows and oranges) do not occur for the same value of the sampling fraction.

3 CONCLUSION

Based on the simulation studies made, we show the inexistence of a single optimum sampling fraction, which would be the best for every single estimation. We also show the inexistence of optimum fractions for the sample mean and variance, since their pattern change for different amount of resamples. However, we noticed the existence of an optimum band for standard error estimators, region that extends from 37.5% to 200% (inclusive).
We have then that for standard error estimation the optimum sampling fraction is smaller than 100%, which translates to a reduction of computing costs compared to higher fractions.

REFERENCES

DAVISON, A. C.; HINKLEY, D. V. Bootstrap methods and their application (Vol. 1). Cambridge: Cambridge University Press, 1997.

EFRON, B.; TIBSHIRANI, R. An Introduction to the Bootstrap. New York: Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1994.

LEWIS, P. A.W.; ORAV, E. J. Simulation Methodology for Statisticians, Operations Analysts and Engineers, Vol 1. Belmont: Wadsworth, 1989.

R CORE TEAM. R: A language and environment for statistical computing. Version 3.6 URL https://www.R-project.org/