Revista Facultad de Ciencias Universidad Nacional de Colombia, Sede Medellín
V 4 N°2 julio-diciembre de 2015 • ISSN 0121-747X / ISSN-e 2357-5749 • Research Paper
Páginas 74 a 82

BOOTSTRAP-BASED INFERENCE FOR GROUPED DATA[a]

INFERENCIA PARA DATOS AGRUPADOS VÍA BOOTSTRAP

JORGE IVÁN VÉLEZ[b], JUAN CARLOS CORREA[c]

Recibido 19-01-2015, aceptado 12-10-2015, versión final 01-12-2015.
Research Paper

ABSTRACT: Grouped data refers to continuous variables that are partitioned in intervals, not necessarily of the same length, to facilitate its interpretation. Unlike in ungrouped data, estimating simple summary statistics as the mean and mode, or more complex ones as a percentile or the coefficient of variation, is a difficult endeavour in grouped data. When the probability distribution generating the data is unknown, inference in ungrouped data is carried out using parametric or nonparametric resampling methods. However, there are no equivalent methods in the case of grouped data. Here, a bootstrap-based procedure to estimate the parameters of an unknown distribution based on grouped data is proposed, described and illustrated.

KEYWORDS: Bootstrap, estimation, grouped Data.

RESUMEN: Los datos agrupados se refieren a variables continuas que se dividen en intervalos no necesariamente de la misma longitud para facilitar su interpretación. Contrario a lo que ocurre en datos no agrupados, la estimación de simples estadísticos de resumen como la media o la moda, o más complejos como un percentil o el coeficiente de variación, es una tarea difícil en datos agrupados. Cuando no se conoce la distribución de probabilidad que genera los datos, la inferencia en datos no agrupados se realiza utilizando métodos paramétricos o no paramétricos de remuestreo. Sin embargo, no existen métodos equivalentes para datos agrupados. En este documento se propone, describe e ilustra un método basado en bootstrap para estimar los parámetros de una distribución desconocida a partir de datos agrupados.

PALABRAS CLAVE: Bootstrap, datos agrupados, estimación.

[a]Vélez, J. I. & Correa, J. C. (2015). Bootstrap-based inference for grouped data. *Revista de la Facultad de Ciencias*, 4 (**2**), 74–82.

[b]The Arcos-Burgos Group, Department of Genome Sciences, John Curtin School of Medical Research, Australian National University, Canberra, ACT 2601, Australia. Neuroscience Research Group, University of Antioquia, Medellín, Colombia. jorge.velez@anu.edu.au

[c]Associate Professor, Department of Statistics,Universidad Nacional de Colombia, Medellín, Colombia. Research Group in Statistics, Universidad Nacional de Colombia, Medellín, Colombia.

1. INTRODUCTION

Data are present in many shapes and with as many variables and observations as one can imagine. Two of the most frequent presentations of data are (1) the raw format, in which observations (i.e., subjects) are represented as rows and variables as columns; and (2) the frequentist or interval format, also known as grouped data, in which, for a particular variable, the number of observations falling in a particular class interval (or bracket) is given in Table 1.

Table 1: Classical representation of univariate grouped data.

Interval	**Frequency**
$[a_1, b_1)$	n_1
$[a_2, b_2)$	n_2
$\vdots$	$\vdots$
$[a_k, b_k)$	n_k
Total	n

Raw data are faced the most by researchers when applying statistical methods and reporting results. However, there are several situations in which researchers have data in interval format. Some examples of this include the age distribution in a sample of people covered by health insurance (i.e., 25-34, 35-44, 45-54 and >55 years old), the number of hours worked per week (i.e., 0-9, 10-19, 20-29, and >30 hours) by part-time employees, and the time taken to complete a task (i.e., 5-10, 10-15, 15-20 and >20 seconds) of people participating in a psychology experiment. As shown in Table 1, a total of n individuals are sampled from the population and the number of them falling on a specific class interval is registered.

Given a random sample of size n, the main goal in applied statistics is to make inferences about a parameter of interest, say θ. This parameter can be a scalar (i.e., the population mean μ, or the population variance σ^2) or a vector (in a multivariate setting). For illustration purposes, let us consider the simplest case in which a random variable X is measured in a sample of n individuals to obtain the random sample $x = (x_1, x_2, \ldots, x_n)$. Suppose that $f_X(x|\theta)$ is the probability density function of X with parameters $\theta = (\mu, \sigma^2)$, and that $\hat{\theta}$ is its sample estimator such that $\hat{\theta} \to \theta$ as $n \to \infty$ almost surely. When the data is not grouped, to make inference about θ is straightforward and most of the statistical methods already available can be applied without difficulty. For instance, if we were interested in constructing a 95 % confidence interval for the true population mean μ based on a random sample of size n, it would be enough to compute an interval of the form $(\hat{\mu} - k\,\hat{\sigma},\ \hat{\mu} + k\,\hat{\sigma})$, where $\hat{\mu}$ is the sample estimator of the population mean, k is a constant that depends on the confidence level and the sample size n, and $\hat{\sigma}$ is the sample estimator of the standard deviation σ (in general, $\hat{\mu}$ and $\hat{\sigma}$ are unbiased estimators of the corresponding population

parameters). Similarly, hypothesis testing and other inferential work can be built upon this same principle.

However, when the sample information for X has been grouped (i.e., the random sample is grouped in k mutually exclusive intervals of some sort as those shown in Table 1), to make inference about θ can become difficult. There are several reasons for this. First, calculating the estimators for population parameters can not be carried out in the same way as if the data were ungrouped. Second, little progress has been made by classical frequency statistics on grouping. Today, the major results are that moment-based statistics are inconsistent, but that maximum likelihood, at least in the normal and exponential cases, is consistent and efficient. Furthermore, to make inferences about (not necessarily linear) combinations of the parameters (i.e., the ratio μ/σ) sets more complex challenges (Heitjan, 1989). Thus, if the researcher wanted to estimate θ or functions or combinations of it, the available methods to accomplish this task cannot, in general, be used. Although some advances have recently been made and applied to income data as illustration (Hajargasht et al., 2012), much research remains to be done to establish a general approach that works in most situations. Tables constructed from continuous data (see Table 1) can be thought as histograms that are density estimators. When the original individual data is available, theoretical and practical methods have been proposed for constructing optimal histograms (Scott, 1979; Taylor, 1987; Kanazawa, 1992; Wand, 1997; Scott and Scott, 2008).

Bootstrap is a powerful technique for finding, via resampling, either the sampling distribution, or the error of an estimator (or a function of it), either of which can be difficult to obtain analytically (Efron, 1979). Since its origins, the bootstrap has extensively been studied by many researchers and its application has been extended to all statistical areas. As the sampling distribution, the error of an estimator and confidence intervals constitute an important and direct byproduct of the technique (Efron, 1987; Hinkley, 1988; Efron, 2003; DiCiccio and Efron, 1996; Letson and McCullogh, 1998; Carpenter and Bithell, 2000; Davison, Hinkley and Young, 2003). In the case of grouped data, the bootstrap provides an easy way to calculate the sampling distribution of $\hat{\theta}$ or any function of it, say $h(\hat{\theta})$.

In this paper we propose a bootstrap-based methodology to make inference about θ when grouped data is available, and which works well in general situations. The paper is organized as follows. First, we briefly described grouped data and how some summary statistics are calculated. Second, we present bootstrap-based methodology by describing a step-by-step general procedure on which our proposal relies. Third, three examples are given to illustrate the usefulness of our proposal. In order to facilitate the implementation of methodology proposed herein, an implementation in R (R Core Team, 2015) can be obtained from the first author by request.

2. GROUPED DATA FROM AN UNKNOWN PARAMETRIC DISTRIBUTION

Let X be a continuos random variable with unknown probability distribution $f_X(x|\theta)$, where $\theta \in \Omega$ the parameter vector, and suppose that $x = (x_1, x_2, \ldots, x_n)$, a random sample of size n, is drawn from $f_X(x|\theta)$. Furthermore, suppose that x_j, $j = 1, 2, \ldots, n$, is classified in one of k possible mutually exclusive categories, with the ith category defined by the half-open interval $[a_i, b_i)$, and the midpoint of the interval is $m_i = (a_i + b_i)/2$, $i = 2, 1, \ldots, k$. A representation of this set up is presented in Table 1. Note that the number of observations in the ith category is n_i and the sample size is $n = \sum_{i=1}^{k} n_i$. It is also worth mentioning that the width of the intervals need not be the same.

2.1. Location statistics

We consider the case where the researcher does not have the raw data, but a tabulated version of it (as in Table 1), and it is of interest to estimate a location parameter θ. In what follows, we address how to estimate θ when grouped data is available. How to perform inference on θ is discussed in §3.

For grouped data the *mean* and *median* can be calculated as (Pierce, 2014):

$$\bar{x}_G = \frac{\sum_{i=1}^{k} m_i\, n_i}{n}, \qquad \tilde{x}_G = a + \frac{n/2 - F_b}{n_m}\, w, \tag{1}$$

where a is the lower limit of the interval containing the median, F_b is the cumulative frequency of the groups before the median interval, n_m is the frequency of the median group and w is the group width. On the other hand, the *mode* can be calculated as (Pierce, 2014):

$$\mathring{x}_G = a + \frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})}\, w, \tag{2}$$

where a is the lower limit of interval containing the modal group, w is the width of the modal group, and n_{m-1}, n_m and n_{m+1} correspond to the frequency of the group before the modal group, the modal group and the group after the modal group, respectively.

2.2. Dispersion statistics

Just as in ungrouped data, it is also possible to estimate dispersion statistics for grouped data. Some of these statistics include, among others, the variance, the standard deviation and the coefficient of variation.

The sample *variance* is calculated as (On, 2002):

$$s_G^2 = \frac{\sum_{i=1}^{k} n_i m_i^2}{n} - \bar{x}^2 \tag{3}$$

and the sample standard deviation as $s_G = \sqrt{s_G^2}$. It is important to mention that s_G^2 is not a good estimate of $\sigma^2 = \text{Var}[X]$, but $s_G^2 - h^2/12$ is, where h is the width of the intervals for X (see Heitjan (1989) for more details). Last, but not least, the sample coefficient of variation (CV_G) can be calculated as

$$CV_G = \frac{s_G}{\bar{x}_G}. \tag{4}$$

3. BOOTSTRAP-BASED INFERENCE FOR GROUPED DATA

In this section, we propose and describe our bootstrap-based method to make inference about an unknown parameter θ using grouped information as in Table 1. As previously mentioned, the data consist of k mutually exclusive categories and, despite that they can be identified within a set of possible values, individual observations (or values) are not exactly known. A direct implication of this is that a set of different values of a variable can be interpreted as being the same just because they fall in the same class interval. However, if the set of values fall in the same group it does not imply, by any means, that they are identical. Similarly, if in a two-dimensional scale two observations lie in the same category, it does not mean they are identical, but that their dimensions were found to lie in the same bracket for both scales (Heitjan, 1989).

Provided that $f_X(x|\theta)$ is unknown, we propose the following algorithm to perform the bootstrap-based inference about θ or a function of it, when grouped data is available:

1. Draw a random sample of size n from a multinomial distribution with probabilities n_i/n, where n_i is the number of observations in the ith interval ($i = 1, 2, \ldots, k$; see Table 1 for more details).

2. Let $n_1^{(b)}, n_2^{(b)}, \ldots, n_k^{(b)}$ be the bth bootstrap frequency table ($b = 1, \ldots, B$). For the ith half-open interval $[a_i, b_i)$, generate a random sample of size $n_i^{(b)}$ from a uniform distribution $U(a_i, b_i)$. Denote this sample as $x^{(b)} = \left(x_1^{(b)}, x_2^{(b)}, \ldots, x_n^{(b)}\right)$, $i = 1, 2, \ldots, k; b = 1, 2, \ldots, B$.

3. For the bth bootstrap sample $x^{(b)}$, compute the test statistic $T(x^{(b)})$, $b = 1, 2, \ldots, B$. Note that the test statistics $T_1, T_2, \ldots, T_B$ will be available for further analyses. Here, T_b refers to the test statistic calculated on $x^{(b)}$.

4. Construct the sampling distribution of $T(\cdot)$ based on $T_1, T_2, \ldots, T_B$.

The algorithm described above could be seen as a three-step sampling strategy. In the first step, a *new* sample size is generated for each of the k mutually exclusive categories (or intervals) in

which the data has been broken down. Secondly, based on those sample sizes, we randomly generate observations within each bracket. Last but not least, a test statistic $T(\cdot)$ is calculated for each complete bootstrap sample to construct its sampling distribution. As a by product of this latter step, inferences about θ or a function of it, $h(\theta)$, can be made.

It is worth mentioning that when $f_X(x|\theta)$ is known (i.e., it can be assumed that observed table is generated from a known theoretical parametric model), both the estimation and inferential problems become an estimation problem where all observations are doubled-censored. For more on this topic, we encourage the reader to review the work by Zhan and Wellner (1995).

4. EXAMPLES

In this section we present three examples to illustrate the usefulness of our proposed method.

Example 1. We collected the weight, in grams, of $n = 159$ coins of COP$100 currently circulating in Colombia as part of a random academic experiment. The weights categorised in seven equally spaced class intervals are presented as follows:

Table 2: Weight of COP$100 coins (in grams).

Interval	Frequency
5.00–5.15	2
5.15–5.20	7
5.20–5.25	29
5.25–5.30	60
5.30–5.35	50
5.35–5.40	9
5.40–5.45	2

Let X be the the weight of a coin in grams, and suppose we are interested in constructing a 95 % confidence interval for the mean (μ), standard deviation (σ) and the coefficient of variation (CV). Following the algorithm described in §3, we generated $B = 10,000$ bootstrap samples from a multinomial distribution of size $n = 159$ where the probabilities of all cells were given by $\hat{\pi} = (2/159, 7/159, 29/159, 60/159, 50/159, 9/159, 2/159)$. Further, a sample of size n_i from a uniform distribution $U(a_i, b_i)$ was drawn ($i = 1, 2, \ldots, 7$). Observe that, at the end of this process, a total of B random samples of size $n = 159$ are generated from $f_X(x|\theta)$. The corresponding 95 % confidence intervals are $\mu \in (5{,}273, 5{,}291)$, $\sigma \in (0{,}048, 0{,}065)$ and $\sigma^2 \in (0{,}009, 0{,}012)$.

Example 2. Harrell and Davis (1982) proposed a distribution-free estimator of the median given by

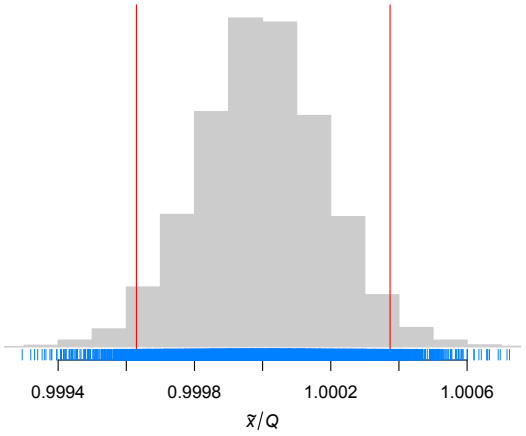

Figure 1: Sampling distribution of $\tilde{x}/Q$ for the coin data presented in Table 2. The red and blue vertical lines correspond to the 95 % confidence interval and the individual values, respectively.

$$Q = \sum_{i=1}^{n} W_{n,i}\, X_{(i)}, \qquad W_{n,i} = \frac{\Gamma(n+1)}{\Gamma\left(\frac{n+1}{2}\right)^2} \int_{(i-1)/n}^{i/n} [z\,(1-z)]^{(n-1)/2}\, dz. \tag{5}$$

Vélez and Correa (2014) have recently shown that Q has a lower mean squared error than the classical estimator of the median when the sample size is small (i.e., $n < 50$). Observe that, with grouped data, the calculation of the Q statistic is not straightforward. However, using our proposed method this calculation becomes simple. Furthermore, it is also possible to derive the sampling distribution of more complex statistics such as $\tilde{x}/Q$, where $\tilde{x}$ is the sample median. Using the data in Table 2, the sampling distribution of $\tilde{\mu}/Q$ was obtained (see Figure 1). It follows that a 95 % confidence interval for $\tilde{\mu}/Q$ is (0.9996, 1.0003). Similarly, $\tilde{\mu} \in (5{,}276, 5{,}288)$ and $Q \in (5{,}278, 5{,}288)$.

Example 3. For ungrouped data, the sample kurtosis can be calculated as

$$\hat{\beta}_2 = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^4}{\{\sum_{i=1}^{n} (x_i - \bar{x})^2\}^2} - 3 \tag{6}$$

and similar expressions can be found for grouped data. Let β_2^G be the equivalent kurtosis coefficient for grouped data. Using the data in Table 2 and our bootstrap-based approach, $\hat{\beta}_2^G = 1{,}716$ and the corresponding 95 % confidence interval is $(0{,}109, 3{,}393)$.

5. DISCUSSION

When data about a continuous variable is presented in form of a histogram or table, the estimation of certain characteristic of interest could be a difficult task. In this paper, we have proposed a

bootstrap-based methodology that is easy to interpret and implement, and computationally feasible. Three examples are presented to illustrate the usefulness of our method in situations often encountered by data analysts, and in which it would have been almost impossible to construct confidence intervals for the parameters of interest otherwise.

Future topics of research might include the construction of other types of confidence intervals and the natural extension of our methodology to bivariate or p-variate grouped data. Another topic worth looking at is the comparison between the results obtained with the method proposed herein, and those obtained when the data is not grouped. One alternative to tackle this would be to control the sample size n, generate continuous data from a known distribution with specific parameters, choose a number of categories k, estimate a statistic of interest with both the grouped and ungrouped data, and compare both values (i.e., using the mean squared error). By doing this, it would not only be possible to evaluate the performance of the bootstrap-based method proposed in this paper and those already established (i.e., central limit theorem [CLT]), but also to determine the optimal number of classes, k_0, needed to obtain at least comparable results between our method and the CLT.

Acknowledgements. We are grateful to Dr. Fernando Marmolejo-Ramos from the School of Psychology, University of Stockholm, Sweden for critical reading of this manuscript; and to three anonymous reviewers for their comments and suggestions. JIV was supported by the Eccles Scholarship in Medical Sciences, the Fenner Merit Scholarship and the Australian National University (ANU) High Degree Research Scholarship. The first author thanks Dr. Mauricio Arcos-Burgos from ANU for his support.

Authors' contributions. JCC conceived the study. JIV analysed the data and wrote the paper.

Conflict of interest. The authors declare that they have no competing interests.

Computational details. R code implementing the method described here is available from the authors by request.

References

Carpenter, J. & Bithell, J. (2000), Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians, *Statistics in Medicine*, 19(**9**), 1141–1164.

Davison, A. C.; Hinkley, D. V. & Young, G. A. (2003), Recent Developments in Bootstrap Methodology, *Statistical Science*, 18(**2**), 141–157.

DiCiccio, T. J. & Efron, B. (1996), Bootstrap Confidence Intervals, *Statistical Science*, 11 (**3**), 189–228.

Efron, B. (1979), Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, 7 (**1**), 1–27.

Efron, B. (1987), Better Bootstrap Confidence Intervals, *Journal of the American Statistical Association*, 82 (**397**), 171–185.

Efron, B. (2003), Second Thoughts on the Bootstrap, *Statistical Science*, 18(**2**), 135–140.

Hajargasht, G.; Griffiths, W. E.; Brice, J. & Rao, D. P. & Chotikapanich, D. (2012), Inference for income distributions using grouped data, *Journal of Business & Economic Statistics*, 30(**4**), 563–575.

Harrell, F. E. & Davis, C. E. (1982), A New Distribution-Free Quantile Estimator, *Biometrika*, 69(**3**), 635–640.

Heitjan, D. F. (1989), Inference from grouped continuous data: A review, *Statistical Science*, 4(**2**), 164–179.

Hinkley, D. V. (1988), Bootstrap Methods, *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(**3**), 321–337.

Kanazawa, Y. (1992), An Optimal Variable Cell Histogram Based on the Sample Spacings, *The Annals of Statistics*, 20 (**1**), 291–304.

Letson, D. & McCullogh, B. D. (1998), Better Confidence Intervals: The Double Bootstrap with No Pivot, *American Journal of Agricultural Economics*, 80(**3**), 552–559.

On, C. W. (2002), Mean, Variance and Standard Deviation for Grouped Data, http://www.angelfire.com/blues/michaelyang/ive/dms/chapter_05/5_6_StaDev.html. Accessed: 2015-11-18.

Pierce, R. (2014), Mean, Median and Mode from Grouped Frequencies, https://www.mathsisfun.com/data/frequency-grouped-mean-median-mode.html. Accessed: 2015-11-18.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Scott, D. W. (1979), On Optimal and Data-Based Histograms, *Biometrika*, 66 (**3**), 605–610.

Scott, D. W. & Scott, W. R. (2008), Smoothed Histograms for Frequency Data on Irregular Intervals, *The American Statistician*, 62(**3**), 256–261.

Taylor, C. C. (1987), Akaike's Information Criterion and the Histogram, *Biometrika*, 74(**3**), 636–639.

Vélez, J. I. & Correa, J. C. (2014), Should we think of a different Median estimator?, *Revista Comunicaciones en Estadística*, 7(**2**), 1–8.

Wand, M. P. (1997), Data-based Choice of Histogram Bin Width, *The American Statistician*, 51(**1**), 59–64.

Zhan, Y. & Wellner, J. A. (1995), Double censoring: characterization and computation of the nonparametric maximum likelihood estimator, Technical report.