ELSEVIER

Computers in Biology and Medicine 38 (2008) 313–328

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

The proportionator: Unbiased stereological estimation using biased automatic image analysis and non-uniform probability proportional to size sampling

J.E. Gardi*, J.R. Nyengaard, H.J.G. Gundersen

Stereology and Electron Microscopy Research Laboratory and MIND Center, University of Aarhus, Ole Worms Allé 1185, DK - 8000, Aarhus C, Denmark

Received 30 April 2007; accepted 13 November 2007

Abstract

The proportionator is a novel and radically different approach to sampling with microscopes based on the well-known statistical theory (probability proportional to size—PPS sampling). It uses automatic image analysis, with a large range of options, to assign to every field of view in the section a weight proportional to some characteristic of the structure under study. A typical and very simple example, examined here, is the amount of color characteristic for the structure, marked with a stain with known properties. The color may be specific or not. In the recorded list of weights in all fields, the desired number of fields is sampled automatically with probability proportional to the weight and presented to the expert observer. Using any known stereological probe and estimator, the correct count in these fields leads to a simple, unbiased estimate of the total amount of structure in the sections examined, which in turn leads to any of the known stereological estimates including size distributions and spatial distributions. The unbiasedness is not a function of the assumed relation between the weight and the structure, which is in practice always a biased relation from a stereological (integral geometric) point of view. The efficiency of the proportionator depends, however, directly on this relation to be positive. The sampling and estimation procedure is simulated in sections with characteristics and various kinds of noises in possibly realistic ranges. In all cases examined, the proportionator is 2–15-fold more efficient than the common systematic, uniformly random sampling. The simulations also indicate that the lack of a simple predictor of the coefficient of error (CE) due to field-to-field variation is a more severe problem for uniform sampling strategies than anticipated. Because of its entirely different sampling strategy, based on known but non-uniform sampling probabilities, the proportionator for the first time allows the real CE at the section level to be automatically estimated (not just predicted), unbiased—for all estimators and at no extra cost to the user.

Keywords: Automatic image analysis; PPS sampling; Section quantization; Simple random; Smooth fractionator; Specific stains; Stereology; Systematic uniform random; Tissue inhomogeneity

1. Introduction

In most available computer aided systems for stereological purposes, the fields of view (FOVs) to be examined by the user are selected using systematic, uniformly random sampling (SURS). The user first delineates the section or the relevant part of it, and the computer then samples a specified fraction of all FOVs roughly equidistantly spaced. The sampling scheme is sometimes referred to as 'meander sampling' (actually referring to one of the ways of implementing it manually).

The essence is that FOVs are sampled with a predetermined, constant probability, irrespective of the content of the fields (there may be no tissue present in some fields, for example). Moreover, in inhomogeneous tissue, many fields contain none or very few positive events and a large number of fields must therefore be examined to obtain a reasonable precision. In the extreme case of rare events, the examination of a very large number of empty fields is almost the whole workload and the resulting precision may still be very low.

It is a consequence of such a sampling scheme that any inhomogeneity in the tissue is noise, an unwanted characteristic of the tissue that, without exception, means more work (in order to obtain a given precision). For decades, numerous attempts have therefore been made to improve the performance using automatic image analysis, generally without any success

* Corresponding author. Tel.: +45 8942 2945; fax: +45 8942 2952.
E-mail address: Jonathan.Gardi@ki.au.dk (J.E. Gardi).

0010-4825/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2007.11.002

because of the low and varying contrast and almost insurmountable difficulties in defining the relevant image segmentation in biological materials.

The sampling paradigm presented here is different. It relies on the fact that all staining methods provide the structure of interest with some particular stain. The stain may occasionally be a specific antibody, but is often of low specificity like the standard hematoxylin–eosin, for example. With that stain, all cell nuclei are varying shades of blue and many other components are reddish. The specificity in counting cell nuclei of a well- defined cell type, for example, is provided by the expert observer who recognizes the cell under study using texture, configuration, neighbor relations and many other aspects that vary from one cell type to another.

Nevertheless, any FOV without blue stain cannot contain the cell nucleus under study (the field may be outside the section or in a region that happens not to contain any cells). On the contrary, a field with much blue may contain many cells of interest. Crudely: no blue, no interest, much blue, much interest.

This is the idea of proportionator sampling. Initially, the computer is used for automatically collecting some relevant information about all fields in the section. Using some arbitrary, predefined algorithm the amount of information is 'measured' (the total amount of blue, for example). The computer then selects a predetermined number of the fields, each with a probability strictly proportional to this amount (hence the name of the sampling paradigm). In the selected fields the expert user then makes the specific, 'correct' count of cell nuclei using all the usual clues, as outlined above—and the disector [1], physical or optical. Finally, the correct count in fields sampled with a known probability provides an unbiased estimate of the total number of cell nuclei in the section.

Note that under this sampling scheme any inhomogeneity of the tissue (with respect to the selected characteristic) becomes a signal that is used for making the sampling more efficient, the opposite of the ordinary situation described above.

The present report is an explorative study of the proportionator and a preliminary report of its performance: the genuine test of a sampling strategy is evidently to study it under realistic circumstances in real tissue. The proportionator is based on an unequivocally unbiased principle, but its efficiency under all kinds of problems and real distributions of structures in biological material is impossible to predict. We have therefore made this simulation study, with all relevant details as close to reality as possible, in order to test the strategy's robustness when pushing the envelope of the simulation in various directions.

The mathematical basis of proportionator sampling is first presented, then the simulation is briefly outlined and finally the simulation results are presented and discussed.

2. The proportionator: sampling fields with probability proportional to size

This is a well-known statistical sampling technique often used in survey sampling [2]. Among statisticians, it is generally referred to as probability proportional to size—PPS sampling. 'Size' is quite abstract, it may be any feature or characteristic that may be quantitated, at least crudely, and which has some positive relation or association to the objects under study. We shall use the amount z_i of a specific color in each FOV as an example; it is also used in the simulation.

z_i must be known for all fields in the whole section, for which reason it evidently must be obtainable automatically. The sum over all N fields is denoted by Z. The N fields are listed in any arbitrary order and the accumulated sum Fz is computed for this ordering.

The sample size wanted is n, and the quantity Z/n serves as a sampling period under SURS, see Fig. 1. In the ordered set of all fields, sampling is uniform in the accumulated sum Fz, the ordinate of Fig. 1. Since Fz is a step function, any number between 0 and Z uniquely identifies an FOV (with known coordinates) among all fields. First a random start is generated between 0 and Z/n. The remaining fields to be sampled are then identified by adding Z/n to the random sampling point of the previous selection. In short, this is ordinary SURS, like taking every seventh section, except that sampling is not among integer-indexed physical items but in the real-valued function Fz. As an unusual consequence, the sample size is a fixed constant (in ordinary SURS the sample size is a random variable).

One would usually study several sections from the same specimen. Optimally, they should be studied as one assembly, but it depends on the technical possibilities (more sections on one glass slide, several slides on the microscope stage, adequate software control of the microscope). If they can only be studied separately, one should aim at roughly a constant sampling period Z/n, thereby sampling most cells from the sections withth most of the indicative color. The adequate sampling period for obtaining a predetermined precision has to be determined in the pilot study, just like in SURS where one has to figure out the adequate sampling distances (step lengths) prior to making the serious observations.

The selected FOVs are presented to the user. For each field, the user enters the correct stereological count, x_i, and the unbiased estimator of the total content X in the section is simply

$$X := \sum_i^n \frac{x_i}{z_i/(Z/n)} = \frac{Z}{n}\sum_i^n \frac{x_i}{z_i}. \tag{1}$$

This is an instance of the general, unbiased Horvitz–Thompson estimator of a population total [3]:

$$\text{PopulationTotal} := \sum^n \frac{\text{Item Content}}{\text{ItemSamplingProbability}}. \tag{2}$$

Note that the unbiasedness of the estimator has no relation to the fact that the amount of color in a field is clearly biased information with respect to the correct nuclear count in 3D according to the disector counting rule—or with respect to any other stereological estimator. At the time of sampling, the number, $z_i = 0.123$, for example, assigned to the field because it is the sum of particularly colored pixels, just guarantees that particular field has a known and fixed sampling probability under proportionator sampling: $\text{probability}_i = z_i/(Z/n) = 0.123/(\frac{200}{20}) = 0.0123$ for $n = 20$ and $Z = 200$. In other words, once it is recorded, z_i is just a fixed number that controls a correct sampling and