Published in METRON - International Journal of Statistics 69, issue 1, 45-65, 2011
which should be used for any reference to this work

MATTI LANGEL – YVES TILLÉ

Corrado Gini, a pioneer in balanced sampling and inequality theory

Summary - This paper attempts to make the link between two of Corrado Gini's contributions to statistics: the famous inequality measure that bears his name and his work in the early days of balanced sampling. Some important notions of the history of sampling such as representativeness, randomness, and purposive selection are clarified before balanced sampling is introduced. The Gini index is described, as well as its estimation and variance estimation in the sampling framework. Finally, theoretical grounds and some simulations on real data show how some well used auxiliary information and balanced sampling can enhance the accuracy of the estimation of the Gini index.

Key Words - Gini; Inequality; Balanced sampling; Representativeness; Linearization; Purposive selection.

1. Introduction

Undoubtedly, the most famous contribution of Corrado Gini to the field of statistics is his work on inequality measures, prominently the Gini index, which is still today the leading inequality measure. It is used in various domains outside statistics such as economics, sociology, demography and is a strong political tool. A probably less known contribution of Gini is to be found in the field of survey statistics. Indeed, in the 1920s, Gini took part in the committee that advocated the use of samples in official statistics. Moreover, some years later he sampled Italian circumscriptions in a way that is often referred to as the first example of a balanced sample. The goal of this article is to provide an understanding of the evolution of the notion of balanced sampling across the twentieth century and how this notion has participated in the debate between random and purposive selection methods. Eventually, we propose to link both of the above contributions of Corrado Gini and show how some well used auxiliary information and balanced sampling can enhance the accuracy of the estimation of the Gini index.

The next section is dedicated to a short overview of the history of sampling and particularly of the debate on purposive and random methods. In Section 3, the controversial notion of representativeness is widely discussed. Balanced sampling and Gini's contribution to the topic are considered in Section 4. Sections 5 and 6 define the Gini index, its expression in the sampling framework as well as the estimation of its sampling variance. The two further sections are dedicated to the linkage between the Gini index and balanced sampling: Section 7 shows how balanced sampling can enhance the accuracy of the estimation of the Gini index by means of a sample, followed by a confirmatory simulation study on real data operated and discussed in Section 8. Finally, the last section of the paper is dedicated to concluding remarks.

2. The early stages of sampling

2.1. Acceptance in the scientific community

The common use of sampling in statistics is recent. Many papers report the historical background of sampling in statistics (for example Hansen *et al.*, 1985; Hansen, 1987; Fienberg and Tanur, 1995, 1996). First, its struggle to be accepted as a valid method in a field that is accustomed to full-coverage methods (census); Second, the confrontation between the two different paradigms inside the field, random selection and purposive selection.

A notorious and early proposition of the use of a sample instead of a full census was done by A.N. Kiaer (1896, 1899, 1903, 1905) at the 5th International Statistical Institute (ISI) meeting in Bern in 1895. Sampling methods would be more widely accepted thirty years later, in 1925, when some results are presented at the 16th ISI Sessions in Rome. These results are those of the Reports on the representative method in statistics, proposed by A. Jansen and a committee of five other statisticians (Jensen, 1926). Corrado Gini is one of them.

The history of sampling theory has been driven by several concepts, often associated to unclear meanings. There is frequent confusion and misunderstanding between the notions of representativeness, purposive selection, random selection or balanced samples. These concepts, as well as their evolution across the history of survey methods, are discussed in the following sections.

2.2. Purposive and random selection

In addition to eventually giving credit to the idea of sampling, the report by Jensen (1926) already opposes two methods of sampling: purposive selection and random selection. This separation has been the vector of a large amount of researches on sampling.

When speaking of random selection it is firstly of importance to distinguish between the population of interest and the sampling frame. The sampling frame is the list of units from which the sample is to be drawn. For example, if a survey on male adults of a given country (the population of interest) is to be done, a list of all male adults in the country has to be generated before sampling can be effective. Ideally, the sampling frame corresponds fully to the population of interest, but for various reasons under-coverage (units that are in the population but not in the frame) or over-coverage (units that should not be in the frame) is possible.

The selection process is referred to as random when all units in the sampling frame have a non-null probability of being selected in the sample and that this inclusion probability can be precisely established. Moreover, the scheme is such that every possible subset (*e.g.* sample) s of the population U has a probability of selection, denoted $p(s)$. Unlike in some purposive methods, the interviewer does not take any part in the selection process of the sample, which is operated by an algorithm. Based on mathematical validation and allowing for the construction of confidence intervals, random sampling is soon preferred in the scientific community, as well as in official statistics. However, purposive selection is still used nowadays, mostly via quota sampling methods.

Purposive selection (or non-probability sampling), regroups any sampling method in which the inclusion probabilities are not known or in which some units have no chance to be selected in the sample. These methods are often used by private polling organizations because these organizations generally do not have access to the sampling frame from which a sample can be drawn (a census for example) and also because the total cost of the survey can be lowered when using non-probability methods.

3. Representativeness

3.1. A polysemic term

The idea and concept of *representativeness* was already used in Kiaer's work (Kiaer, 1896, 1899, 1903, 1905). Because the idea of a *representative sample* is reassuring for an uninitiated audience as it provides an illusion of scientific validity, it has been an important notion in sampling ever since. However, the multiplicity of definitions to which it can be associated has been at the core of many debates and misunderstandings in the history of sampling. Thus, the term is much less used in modern survey sampling literature and in our opinion it is a term best to avoid in survey methodology.

Kruskal and Mosteller (1979a,b,c, 1980) have written several survey papers in which they typologize and the different definitions of representativeness in

and out of the field of statistics. They have listed nine different views of the word and illustrated them by numerous excerpts from the literature. From their work, one can also see that the definition of representativeness and the question of purposive or random selection are often directly linked.

3.2. Representativeness of the sample

One point of interest which is seldom clarified is the statistical object to which the quality of representativeness is attributed. It could be the statistical unit to be sampled (a representative individual, a representative district or firm), the sample itself or the strategy used for the sample selection procedure. Indeed, for example, using a method of sampling which supposedly provides a representative sample and analyzing a posteriori the level of representativeness of an accomplished sample are two very different conceptions of the problem. Also considering the question of randomness and representativeness, Hájek (1959, 1981) invokes the term of representative strategy (and not representative sample) when talking about sampling methods like balancing or calibration, a strategy being a pair composed of a sampling design and an estimator. When used hereafter, the term of representativeness is applied to the sample, except where specified.

3.3. Representativeness as a miniature of the population

Although many decisive differences occur between the authors, the underlying idea of representativeness in most cases is that to be considered representative a sample should be a miniature of the population of interest. Obviously, this is a purely theoretical construct, which creates many divergences in its application. Indeed, in order to select a sample which would be a downsized replica of the population, the whole complexity of the population structure has to be known, which is impossible. Generally, only a small aspect of this structure is known, turning representativeness into a much more relative notion. Moreover, the more information one has in the population, the more dissimilar are each of its units. In the end, the only perfectly representative sample would thus be the population itself. Also, as discussed further in this section, the results of Neyman (1934) in optimal stratification show that, in fact, expecting a sample to be a miniature of the population is an erroneous ideal.

In practice, it is therefore only possible to come close to this view on representativeness and many different solutions have been proposed to achieve it, some of which being totally antagonistic. These methods depend obviously on whether purposive selection or random selection is preferred, but also on the different interpretations of representativeness. An example of these divergences is well described by the question of the presence or absence of selective forces, raised by Kruskal and Mosteller (1979a,b,c, 1980) and discussed below.

3.4. Selective forces

It can be advocated that the complete absence of a selective process involving any kind of auxiliary information is a guarantee of representativeness. This idea fosters the use of simple random sampling, because in that framework every unit in the population has the same probability of being selected in the sample, regardless of any of the unit's characteristics. Here, selective forces are viewed as a curb to representativeness, and the definition of the latter is partially confused with randomness. Moreover, the notion of randomness itself seems bounded to simple random sampling, because other designs like stratification or unequal probability sampling rely on selective forces.

On the contrary, other definitions see in selective forces a necessary requirement to guarantee representativeness. For instance, this can be illustrated by the widespread idea that the more a sample presents the same characteristics (for example gender ratio, age groups, mean income) as the population of interest, the more it can be declared representative. For the latter, perfect representativeness is not accessible in practice, but can nonetheless be understood as a miniature of the population and approached by using the available information as controls.

Likewise, some have an intermediate position. More than two decades after his work with Corrado Gini (Gini and Galvani, 1929), which is detailed in later sections of this paper, Galvani (1951) distinguishes three different procedures of sampling: random selection, purposive selection and stratified selection.

In his paper, random selection is what is more often known as simple random sampling. It uses no auxiliary information, and therefore requires no prior knowledge of the population to be sampled. For the author, it is also the procedure which is closest to the idea of reduction of the totality, where the totality would be the whole population and all its characteristics. According to Galvani, this statement does not imply that any achieved sample from a simple random sampling design is effectively representative for all characteristics of the population. It could be therefore argued that Galvani is unclear whether representativeness is an attribute of the accomplished sample or of the method of selection.

Purposive selection, instead, requires some auxiliary knowledge on the whole population which are used to guarantee the representativeness of the sample in relation to these auxiliary variables. However, Galvani notes that, unlike in the random selection, the purposive procedure can by no means be considered representative for characteristics that are not used in the selection process. If we summarize Galvani's point of view, he considers on one side simple random sampling which, as a *method*, is representative for all characteristics in the population, and on the other side, purposive selection which yields a *sample* that can be described as representative for only a closed number of

characteristics. This leads him to oppose absolute representativeness (random selection) and relative representativeness (purposive selection), the latter being described as inferior to the former.

Stratified sampling is coherently considered by Galvani as a random selection procedure which makes use of some knowledge of the heterogeneity of the population regarding some characteristic. The property of "impartiality", or reduction of the totality, can also be credited to the stratified sampling method, with presumably a better accuracy than for simple random sampling. In other words, Galvani thinks that stratification is not a type of selective force which can jeopardize representativeness. He states that random selection methods should be preferred because they satisfy fully to the conditions of representativeness. Moreover, probability theory can be applied to the random case.

3.5. Representativeness as a purpose

In the Kruskal and Mosteller (1979a,b,c, 1980) papers emphasis is not put on one essential question: should representativeness be a purpose for survey sampling? From our point of view, the goal of a survey is to provide good estimations (in terms of bias and variance) of some characteristics of the population by means of a sample. In that framework, representativeness can possibly be a means for achieving good estimations, not a goal in itself.

Probability sampling theory and an adequate use of auxiliary information has shown that using unequal probabilities of inclusion or over-representing some groups of the population can enhance the accuracy of the estimation. In that sense, a sample can be obviously very far from being a miniature of the population and estimate efficiently the characteristics of interest. The famous paper by Neyman (1934) stresses two major assets of probability sampling theory: only probability sampling methods can be theoretically validated because they are the outcome of a random experiment; the results on optimal stratification prove that a representative sample is not optimal in terms of accuracy. The latter result, which can be viewed as counter-intuitive is a major defeat for non-probability sampling, as well as for the idea of representativeness as a purpose in the field of survey sampling.

4. Balanced sampling

4.1. A broad definition

Representativeness and randomness are both major debates of sampling theory and the development of balanced sampling has certainly changed their physiognomy. Balanced sampling, as defined below, has put the conception of

the representative sample as a miniature of the population in practice, first under the purposive selection paradigm, than in the random sampling framework. A balanced sampling strategy is a method of selection which uses auxiliary information at the design phase. Moreover, a sample is said to be balanced if its natural estimators of total on some auxiliary variables X will match (or approximately match) the known population totals of these variables. What is meant here by a natural estimator is an estimator such that the weight of any given unit does not change from a sample to another. The main challenge of balanced sampling is the selection process, because the sample has to be selected with respect to the balancing constraints.

This broad definition is consistent with purposive as well as random methods. A more formal definition for the random case is given in Section 7.2 when the Cube method is introduced. Coherently, balanced sampling is at first classified as a purposive selection method. One of the first known application of balanced sampling, proposed in Italy by Gini and Galvani (1929) has enhanced this idea. Later, it will be shown that *balanced* and *random* are not two mutually exclusive concepts. These findings have modified the definition of randomness in sampling, which in the beginning of sampling theory was mostly reserved to simple random sampling.

4.2. Gini and the premices of balanced sampling

Not long after the ISI Session in Rome in 1925, a new census is to be run in Italy and room has to be made for the new data. The Italian statistical office wants, however, to keep a *representative* sample of the previous census of 1921. To do so, Gini and Galvani (1929) use a method referred to as purposive selection by Neyman (1952), but also a premise of the idea of balanced sampling (Yates, 1960). At that time, Italy is separated into 214 circumscriptions. The authors propose to keep a sample containing all units inside 29 circumscriptions. The 29 circumscriptions are not selected at random, but instead they are selected in order to match the population means of some important variables. Indeed, the authors selected seven control variables and selected the sample of 29 circumscriptions in order for the sample means to be as close as possible to the population means. As a result, they realize that for most other variables that were not included in the balancing procedure, the match between the population mean and the sample mean is very poor.

Both Neyman (1952) and Yates (1960) discuss the paper of Gini and Galvani (1929) to condemn the purposive selection method. In a different way, they both stress out that, from a sampling point of view, the selection of 29 circumscriptions is a small sample of huge units. Moreover, Yates underlines the fact that the reliability of the results is not assessable with the purposive selection method. Neyman recalls that to obtain a reliable sample,“we must rely

on probability theory and work with great numbers" (Neyman, 1952, p. 107). Whereas the total number of people in the balanced sample of Gini and Galvani is very large, it remains a small sample of 29 units from a sampling point of view. Of course, Neyman advocates that the circumscriptions should have been considered as strata, instead of sampling units.

4.3. Balanced sampling towards randomness

Although representativeness and purposive selection are often associated, Yates (1960, p. 39) has stressed that their is no contradiction between the balanced procedures and random sampling. Furthemore, according to Yates, a balanced sample is only satisfactory if it is random. He proposes a method for selecting a random balanced sample. The randomization is done by first drawing a preliminary random sample and then by selecting a further unit. The latter is compared to the first unit selected in the original sample. If the new unit improves the balance, it is kept in the sample in place of the original unit. If not, it is rejected. Then, the second unit in the original sample is compared to another new unit, and so on. This procedure is repeated until the balance is considered satisfactory. An appealing remark from Yates is also that purposive selection balanced sampling lead to the selection of units for which the balanced variables take a value close to the population mean, resulting with a problematic smaller variability in the sample than in the population.

While Galvani (1951) discussed three different kind of sampling (see Section 3.4), a similar classification can be found in an article by Royall and Herson (1973) which compares three kinds of balanced samples. The first category is purposive selection (like in Gini and Galvani's experiment), the second category is random sampling and the third is what the authors call restricted randomization. For the authors, (simple) random sampling provides a balanced sample "*on the average*" (Royall and Herson, 1973, p. 887), but despite the fact that adjustments can be done with post-stratification, the method can produce severely unbalanced samples. This reasoning on balancing and simple random sampling is very close to Galvani's statement on representativeness. We think however, that the argument of simple random sampling being balanced in average is pointless. It seems to be simply another way of saying that the estimators are unbiased under the sampling design. Furthermore, whereas this is true for total or functions of totals, it is not true for other type of statistics such as ratios, inequality indicators or quantiles.

Restricted randomization, on the other hand, is defined here as a selection process which provides an approximately balanced sample without renouncing to randomization. The selection process is as proposed by Yates (1960) and discussed above. Royall and Herson (1973) point out that none of the methods can however guarantee that a balanced sample is also balanced on variables

that are not included in the selection process. Indeed, the balanced sample can mimic the population only to a certain extent, which depends greatly on the availability of the auxiliary information. For the authors, the experiment by Gini and Galvani is a notorious example of an approximately balanced sample which was found out to be unbalanced for numerous other external characteristics.

This discussion on balanced sampling requires some remarks on the variance of the estimator. Indeed, it has been shown (Deville and Tillé, 2005; Fuller, 2009) that the variance under a balanced sampling design can be expressed as the variance of regression residuals. It is thus obvious that the more the auxiliary variables are correlated with the character of interest, the more the variance is reduced. Therefore, even if the sample is not balanced on a particular variable, the variance of the sample mean and total of that variable is nevertheless reduced if a correlation exists between the latter variable and the auxiliary information.

In the last two decades, a lot of research has been conducted on the idea that balanced sampling is not in contradiction with the random sampling framework. Modern methods allow indeed for the selection of balanced samples and at the same time stay consistent with the notion of randomness (Deville, 1988, 1992; Deville *et al.*, 1988; Deville and Tillé, 2004; Hedayat and Majumdar, 1995; Nedyalkova and Tillé, 2009; Tillé and Favre, 2004, 2005). It is therefore clear that the history of balanced sampling has somewhat blurred the initial separation between purposive selection and random sampling.

5. The Gini index

5.1. An inequality measure

Although we have, in this paper, introduced Corrado Gini through his work in balanced sampling, his most famous contribution is undoubtedly the Gini coefficient or Gini index (Gini, 1912, 1914, 1921), an inequality index based on the Lorenz curve (Lorenz, 1905). An impressive amount of literature has been written on the Gini index (for a survey paper see Xu, 2004), which is still nowadays the most commonly used inequality measure. Finite population inference for the Gini index has been the center of countless discussions and papers. Indeed, many finite population expressions of the index co-exist. Moreover, variance estimation is not straightforward, and robustness issues are frequent, especially when working with income data which is known to be generally heavily skewed.

The contribution of Corrado Gini's index to inequality measure is immense. It has been the most widely used inequality measure for now nearly a century.

Its graphical interpretation, its synthetical comprehension (often written as a percentage, where 0% is full equality and 100% perfect inequality) and the fact that it satisfies many properties of the axiomatic approach to inequality (Cowell, 1988; Dalton, 1920; Pigou, 1912; Shorrocks, 1980, 1984) has favored its preeminence inside the field of inequality theory. Although other measures have gain interest more recently, the Theil index for example (Theil, 1967, 1969) because of its subgroup decomposability property, or the Quintile Share Ratio (Eurostat, 2005; Hulliger and Munnich, 2006; Langel and Tillé, 2011), because of its simpler interpretation for non-specialists, the Gini index is still by far the most applied and studied measure of income inequality.

5.2. Continuous case

There is at least three ways of apprehending the Gini index. The first one is based on the Lorenz Curve, which graphically represents the share of total income earned altogether by a given share of income earners, ordered from poorest to richest. For example it is possible that the poorer 75% of the population earn only 25% of the total income. This understanding of the Gini index is the most common because it is graphically straightforward and gives an immediate definition for the continuous case. Indeed, the Gini index is the ratio of the area between the Lorenz Curve of the population of interest and the diagonal (the Lorenz Curve under perfect equality) and the area below the diagonal. The latter area is always equal to 1/2. Therefore, considering:

(1) a continuous and differentiable strictly increasing cumulative distribution function $F(y)$ of income y in $\mathbb{R}^+$, and $f(y)$ its derivative and probability density function,

(2) Q_α, the quantile of order α, such that $F(Q_\alpha) = \alpha$ and the quantile function $Q(\alpha)$, which can be written as the inverse of the cumulative distribution function: $Q(\alpha) = F(\alpha)^{-1}$,

(3) the Lorenz function (or quantile share) $L(\alpha)$, which is the share of total income earned by all the income earners up to quantile α

$$L(\alpha) = \frac{\int_0^{Q_\alpha} uf(u)du}{\int_0^{\infty} uf(u)du}, \tag{5.1}$$

the definition of the Gini index for an infinite population is then:

$$G = 2\left(\frac{1}{2} - \int_0^1 L(\alpha)d\alpha\right). \tag{5.2}$$

5.3. Discrete case

For the purpose of measuring inequality in a finite population, the partial sum appears to be a central tool. We propose hereafter two definitions of the partial sum, which can in either case be understood as the sum of all incomes smaller or equal to a given quantile. To link the discrete and continuous cases, it can be noted that the partial sum in the continuous case would be expressed by the numerator of the Lorenz function (5.1).

Let U define a finite population of size N, and y_k the income (or other characteristic of interest) of unit $k \in U$. The incomes y_k are assumed to be sorted in increasing order such that k also denotes the rank. A natural expression for the partial sum would therefore be:

$$\sum_{k \in U} y_k \mathbb{1}\left[y_k \leq Q_\alpha\right],$$

where Q_α is the quantile of order α and where $\mathbb{1}[A]$ is equal to 1 if A is true and 0 otherwise. Quantile Q_α can be defined in many different ways in the finite population context (see for example Hyndman and Fan, 1996). The other definition below has two good properties: it gets around the above issue of the finite population quantile and it is strictly increasing upon α:

$$Y(\alpha) = \sum_{k \in U} y_k H\left[\alpha N - (k-1)\right], \tag{5.3}$$

where

$$H(x) = \begin{cases} 0 & \text{if} \quad x < 0, \\ x & \text{if} \quad 0 \leq x < 1, \\ 1 & \text{if} \quad x \geq 1, \end{cases} \tag{5.4}$$

is the cumulative distribution function of a uniform distribution in $[0; 1]$. A specific case for (5.3) is the population total, hereafter simply denoted Y:

$$Y(1) = \sum_{k \in U} y_k = Y.$$

The second property leads to an important result for the Lorenz Curve. The Lorenz Curve, which can be simply written by

$$L(\alpha) = \frac{Y(\alpha)}{Y},$$

is indeed strictly increasing and convex in $[0; 1[$. Recalling (5.2), the Gini index for the continuous case, an expression for a finite population is:

$$G = 2\left(\frac{1}{2} - \frac{1}{Y}\int_0^1 Y(\alpha)d\alpha\right). \tag{5.5}$$

With Expression (5.3), the Gini index becomes:

$$G = \frac{2}{YN}\sum_{k\in U} k y_k - \frac{N+1}{N} = \frac{\sum\limits_{k\in U}\sum\limits_{\ell\in U}|y_k - y_\ell|}{2NY}.$$

5.4. Sampling from finite population

The level of inequality of a finite population, a country for example, is most often estimated by means of a sample. National statistic institutes around the world usually work with complex sampling designs which allow for the use of auxiliary information to enhance accuracy of the statistics of interest. For this reason, estimation and variance estimation of the Gini index under complex sampling designs has been studied profusely.

A sample $s \subset U$ of size $n(s)$ is a subset of the population. A random sample S is selected from U by means of a sampling design $p(s) \geq 0$, for all $s \subset U$ and

$$\sum_{s\subset U} p(s) = 1.$$

The inclusion probability π_k is the probability of unit k to be in the sample. Inclusion probabilities are defined by the sampling design and are such that

$$\pi_k = \sum_{s\ni k} p(s), \text{ for all } k \in U.$$

Some sampling designs are of fixed size, *e.g.* $\text{var}[n(s)] = 0$. For these designs, the sample size is simply denoted as n. Also, if the design is of fixed size, then

$$\sum_{k\in U} \pi_k = n.$$

In sampling theory, a classical estimator of total Y is

$$\widehat{Y} = \sum_{k\in S} w_k y_k,$$

where w_k denotes the sampling weight of unit k. The weight can simply be the inverse of inclusion probabilities (the estimator of total is then known as the Horvitz-Thompson estimator) but can also account for calibration and non-response adjustments. A natural estimator of $L(\alpha)$ would then be

$$\widehat{L}(\alpha) = \frac{\widehat{Y}(\alpha)}{\widehat{Y}},$$

where $\widehat{Y}(\alpha)$ is the plug-in estimator of (5.3):

$$\widehat{Y}(\alpha) = \sum_{k \in S} w_k y_k H\left(\frac{\alpha \widehat{N} - \widehat{N}_{k-1}}{w_k}\right),$$

where, H is as defined in (5.4) and $\widehat{N}_k = \sum_{\ell \in S} w_\ell \mathbb{1}(y_\ell \leq y_k)$. An important feature of this estimator is that the cumulative weight $\widehat{N}_k$ is an estimator of the rank of unit k in the population. The estimation of the rank is one of the most sensible issues in the estimation of the Gini index in the sampling framework. If omitted, this issue leads to substantial errors in the estimation of the sampling variance of the index. Note also that $\widehat{N} = \sum_{k \in S} w_k$. An estimator for (5.5) is then defined by:

$$\widehat{G} = \frac{2}{\widehat{N}\widehat{Y}} \sum_{k \in S} w_k \widehat{N}_k y_k - \left(1 + \frac{1}{\widehat{N}\widehat{Y}} \sum_{k \in S} w_k^2 y_k\right) = \frac{\sum\limits_{k \in S} \sum\limits_{\ell \in S} w_k w_\ell \, |y_k - y_l|}{2\widehat{N}\widehat{Y}}.$$

6. Variance estimation for the Gini index

6.1. Linearization: the influence function approach

The Gini index is nearly one century old and has been used in countless empirical applications since then. However, within these applications, it is not uncommon to find only point estimators, lacking variance or standard deviation estimations, which are nevertheless necessary for confidence intervals construction. The question of variance estimation for the Gini index is not trivial and has motivated a great amount of research (for example Hoeffding, 1948; Glasser, 1962; Ogwang, 2000; Sandstrom *et al.*, 1985; Deville, 1996; Dell *et al.*, 2002; Leiten, 2005; Giles, 2004; Berger, 2008).

One of the main methods of variance estimation of complex statistics in sampling from finite population is the linearization technique. Based on Taylor series, the method has been introduced by Woodruff (1971) and has since motivated many publications presenting different approaches. Estimating equations (Binder and Patak, 1994; Binder, 1996; Kovacevic and Binder, 1997), influence functions (Deville, 1999) and the Demnati-Rao approach (Demnati and Rao, 2004) are all approaches which can be, or have been, applied to variance estimation for the Gini index in complex surveys.

A generalized linearization method based on influence functions has been developed by Deville (1999). The method is derived from the influence function

as defined in the field of robust statistics (Hampel *et al.*, 1985). The influence function in Deville (1999) is slightly different from the latter and is defined by

$$IT(M, x) = \lim_{t \to 0} \frac{T(M + t\delta_x) - T(M)}{t},$$

where measure M allocates a unit mass to each x_k, T is a functional associating a real number or a vector to each measure, and δ_x the Dirac measure at x. In the sampling framework, measure M is estimated by $\widehat{M}$ with mass w_k at each point x_k in the sample. Moreover, the plug-in estimator of functional $T(M)$ is simply $T(\widehat{M})$. The influence function $z_k = IT(M, x_k)$ is a linearized variable of $T(\widehat{M})$ and

$$\frac{T(\widehat{M}) - T(M)}{N^\alpha} \approx \frac{1}{N^\alpha} \left(\sum_{k \in S} w_k z_k - \sum_{k \in U} z_k \right).$$

The variance of $T(\widehat{M})$ can thus be approximated by the variance of the total of the linearized variable

$$\text{var}\left(T(\widehat{M})\right) \approx \text{var}\left(\sum_{k \in S} w_k z_k \right).$$

Thus, the variance of a complex statistic can be estimated under any sampling design for which the expression of the variance of a total is available. Generally, however, the computation of the influence function requires information on the whole population, not only on the sample. Therefore, plug-in estimators $\hat{z}_k$ are used instead of z_k.

6.2. Linearization of the Gini index

The Gini index is not an easy expression to linearize. Solutions based on Taylor linearization resulted in a greatly over-estimated variance (Nygard and Sandstrom, 1985; Sandstrom *et al.*, 1985, 1988). Some further results use estimating equations (Kovacevic and Binder, 1997) or introduce the influence function approach (Monti, 1991; Deville, 1999). Applying the latter approach, a linearized variable for the Gini index is

$$z_k = \frac{1}{NY} \left[2N_k(y_k - \overline{Y}_k) + Y - Ny_k - G(Y + y_k N) \right]. \tag{6.1}$$

As emphasized previously, this expression involves unavailable information at the population level. It can be substituted by the plug-in estimator

$$\widehat{z}_k = \frac{1}{\widehat{N}\widehat{Y}} \left[2\widehat{N}_k(y_k - \widehat{\overline{Y}}_k) + \widehat{Y} - \widehat{N}y_k - \widehat{G}(\widehat{Y} + y_k\widehat{N}) \right],$$

with

$$\widehat{\overline{Y}}_k = \frac{\sum_{\ell \in S} w_\ell y_\ell \mathbb{1}(y_\ell \leq y_k)}{\widehat{N}_k}.$$

The linearization approach has proved its efficiency for variance estimation in several simulation studies in the literature (Osier, 2006, 2009; Dell *et al.*, 2002; Berger, 2008).

7. Balanced sampling and the Gini index

7.1. Linking two of Gini's main contributions

The estimation of the Gini index by means of a sample is of importance to provide reliable information on the level of income inequality in a population. However, the Gini index is known to be very sensitive to high incomes and the stability of the estimator is therefore very dependent upon the presence of extreme values in the sample. Moreover, when the income distribution contains outliers (very high incomes), the sampling distribution of the Gini index becomes skewed, which creates difficulties when constructing confidence intervals, even if the variance of the index is correctly estimated. In this setup, the choice of the sampling design is crucial. Balanced sampling has proved to make good use of auxiliary information when available. However, the method can be only applied to population totals of auxiliary variables. The idea of Lesage (2008) is to propose methods of balanced sampling for non-linear statistics.

7.2. The Cube Method

Deville and Tillé (2004, 2005) have proposed a general procedure to select a balanced sample called the Cube method. The algorithm is non-rejective and selects a sample with respect to the balancing constraints

$$\sum_{k \in S} \frac{x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj},$$

for all auxiliary variable $j = 1, \ldots, p$. As such, balancing is operated by Horvitz-Thompson estimators of totals for the auxiliary variables. Moreover, the statistic of interest is usually also a total (or a function of a total), say $\widehat{Y}$. If the character of interest y_k is well explained by the auxiliary information x_{kj}, the variance of $\widehat{Y}$ is supposedly small.

7.3. Non-linear balancing constraints

In the case of estimating a strongly non-linear function like the Gini index, the use of available auxiliary information is also desired. It can happen that when estimating the Gini index for a population at a time t, the incomes of a previous period in time $t - \delta t$ are known in the population. Denoting x_k, the income of a previous year for unit k, a simple use of this auxiliary information at the design phase would be to balance the sample on the total $X = \sum_{k \in U} x_k$. However, being eventually interested in the estimation of the Gini index, it seems more favorable to balance, not on the total X, but on the Gini index of the x_k's. Lesage (2008) uses for balanced sampling a well-known idea of variance estimation for complex statistics. Indeed, the sampling variance of a statistic $\hat{\theta}$ is easily expressed under a variety of complex sampling designs as long as $\hat{\theta}$ is a total or a function of totals. If instead, $\hat{\theta}$ is a non-linear statistic, the popular approach of linearization consists in bringing the problem back to one of a total, by using the linearized variable.

In the balancing procedure, a similar trick can be used: to balance on a non-linear statistic, the statistic of interest can be linearized and the total of the linearized variable used as a balancing constraint. Moreover, this method does not require any additional computing tools as for standard balanced sampling.

8. Simulation studies

A simulation study has been operated on real household taxable income data from the canton of Neuchâtel, Switzerland to show the relevance of using non-linear balancing constraints. Complete data is available for a population of $N = 82'489$ households for two consecutive years, 2005 and 2006. The goal of this simulation study is to estimate the Gini index in 2006 by means of a sample and evaluate if balancing procedures are able to reduce the variance of the estimator. The character of interest, denoted y_k, is the income of year 2006 for household k. The auxiliary information is available at the population level and expressed as follows:

- π_k: the inclusion probability of unit k,
- x_{k1}: the income of year 2005 for unit k,
- x_{k2}: the linearized variable for the Gini index, as expressed in (6.1), of year 2005 for unit k.

The results for four different sets of simulations are compared. All four simulations consist in drawing 1000 samples of fixed size $n = 5000$. Equal inclusion probabilities $\pi_k = n/N$, for all $k \in U$, are used across all the simulations. The sampling strategies concerning balancing constraints are detailed in Table 1.

Balanced samples have been selected using the algorithm of the Cube method (see Section 7.2). With the equal inclusion probabilities used here, simulation 1 is equivalent to a simple random sampling design without replacement.

TABLE 1: *Simulations: descriptions of the sampling strategies.*

Simulation	Inclusion probabilities	Balancing variables
simulation 1	π_k.	π_k.
simulation 2	π_k.	π_k, x_{k1}.
simulation 3	π_k.	π_k, x_{k2}.
simulation 4	π_k.	π_k, x_{k1}, x_{k2}.

The results of the simulation study is summarized in Table 2. The second column presents the relative bias of the estimated Gini index $\widehat{G}$ computed over all 1000 samples in each simulation set expressed as

$$RB\left(\widehat{G}\right) = \frac{E_{\text{sim}}\left(\widehat{G}\right) - G}{G},$$

where $E_{\text{sim}}\left(\widehat{G}\right)$ is the mean of the estimated Gini index over the 1000 samples and G is the true value of the Gini index in the population. The last column describes the gain in terms of sampling variance of simulations 2, 3 and 4 relatively to simple random sampling (simulation 1).

TABLE 2: *Simulation results.*

Simulation	$RB\left(\widehat{G}\right)$	$\text{var}_{\text{sim}}\left(\widehat{G}\right)/\text{var}_{\text{sim1}}\left(\widehat{G}\right)$
simulation 1	−0.009%	1.000
simulation 2	0.036%	0.773
simulation 3	0.000%	0.472
simulation 4	−0.020%	0.451

The four Monte Carlo simulations show that the bias of $\widehat{G}$ is negligible. The sampling variance of the estimator is lowered for all balanced designs in comparison with the simple random sampling design (simulation 1). While balancing on x_{k1}, the income of the previous year, clearly has an effect on the variance (simulation 2), the most important result here is that balancing on x_{k2}, the linearized variable of the Gini index of the previous year yields a far better improvement (simulation 3). The same initial auxiliary information is available in both simulations 2 and 3, but the use of the linearized variable presented in Section 6 instead of the plain incomes of year 2005 gives a definitely more stable estimator. In our case, balancing on x_{k2} results in a sampling variance

which is more than twice smaller than with simple random sampling. Finally, balancing on both x_{k1} and x_{k2} (simulation 4) gives the best result but essentially shows that when x_{k2} is used, also adding x_{k1} as a balancing constraint does not bring much gain in terms of variance.

9. Conclusion

In this paper, we have reviewed two of Corrado Gini's main contributions and how they have participated in the development of their respective field. For a start we have discussed the ambiguous notions of representativeness, randomness and balanced sampling in order to get the article of Gini and Galvani (1929) back in its perspective and show how it has participated in the debates that have resulted in modern survey sampling theory.

Secondly, we have presented the Gini inequality index and its application in the sampling framework. The linearization of the Gini index through the influence function method is also introduced for two distinct goals. The first one is variance estimation, which is briefly discussed. The second one is for balanced sampling, which is of particular interest in this paper. Indeed, we have shown how balanced sampling and the use of a linearized variable as a balancing constraint can improve estimation. With the help of a simulation study on real data we have shown that, if the information is available, balancing on the income of a previous year when estimating the Gini index has a positive impact on the stability of the estimator. However, we have also shown that this auxiliary information can be used in a much better way, that is using the linearized Gini index of a previous year as balancing constraints instead of the plain incomes.

Acknowledgments

The authors are grateful to the Office Cantonal de la Statistique (Canton de Neuchâtel, Switzerland) and especially to Gérard Geiser for the dataset. This research is supported by the Swiss National Science Foundation (grant no. 200021-121604).

References

Berger, Y. G. (2008) A note on asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient, *Journal of Official Statistics*, 24, 541-555.

Binder, D. A. (1996) Linearization methods for single phase and two-phase samples: a cookbook approach, *Survey Methodology*, 22, 17-22.

Binder, D. A. and Patak, Z. (1994) Use of estimating functions for estimation from complex surveys, *Journal of the American Statistical Association*, 89, 1035-1043.

Cowell, F. A. (1988) Inequality decomposition: Three bad measures, *Bulletin of Economic Research*, 40, 309-311.

DALTON, H. (1920) The measurement of the inequality of incomes, *The Economic Journal*, 30, 348-361.

DELL, F., D'HAULTFOEUILLE X., FEVRIER, P. and MASSE, E. (2002) Mise en oeuvre du calcul de variance par linéarisation, *INSEE-Methodes : Actes des Journees de Methodologie Statistique*, 73-104.

DEMNATI, A. and RAO, J. N. K. (2004) Linearization variance estimators for survey data (with discussion), *Survey Methodology*, 30, 17-34.

DEVILLE, J.-C. (1988) *Estimation linéaire et redressement sur informations auxiliaires d'enquêtes par sondage*, In: Essais en l'honneur d'Edmond Malinvaud (Eds. A. Monfort, and J. J. Laffond), Economica, Paris, pp. 915-929.

DEVILLE, J.-C. (1992) *Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information*., In: Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys, Örebro, Sweden.

DEVILLE, J.-C. (1996) Estimation de la variance du coefficient de Gini estimé par sondage, *Actes des journées de Méthodologie Statistique, INSEE*, 69-70-71, 269-288.

DEVILLE, J.-C. (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques, *Survey Methodology*, 25, 193-204.

DEVILLE, J.-C., GROSBRAS, J-M. and ROTH, N. (1988) *Efficient sampling algorithms and balanced sample*, In: COMPSTAT, Proceedings in Computational Statistics, Physica Verlag, Heidelberg, pp. 255-266.

DEVILLE, J.-C. and TILLÉ, Y. (2004) Efficient balanced sampling: The cube method, *Biometrika*, 91, 893–912.

DEVILLE, J.-C. and TILLÉ, Y. (2005) Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, 128, 569–591.

EUROSTAT (2005) *The continuity of indicators during the transition between ECHP and EU-SILC*, Technical report, Working Papers and Studies, Luxembourg: Office for Official Publications of the European Communities.

FIENBERG, S. E. and TANUR, J. M. (1995) Reconsidering Neyman on experimentation and sampling: Controversies and fundamental contributions, *Probability and Mathematical Statistics*, 15, 47–60.

FIENBERG, S. E. and TANUR, J. M. (1996) Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling, *International Statistical Review*, 64, 237–253.

FULLER, W. A. (2009) Some design properties of a rejective sampling procedure, *Biometrika*, 96, 933–944.

GALVANI, L. (1951) Révision critique de certains points de la méthode représentative, *Revue de l'Institut International de Statistique*, 19, 1–12.

GILES, D. E. A. (2004) Calculating a standard error for the Gini coefficient: some further results, *Oxford Bulletin of Economics and Statistics*, 66, 425–433.

GINI, C. (1912) *Variabilità e Mutabilità*, Tipografia di Paolo Cuppin, Bologna.

GINI, C. (1914) Sulla misura della concentrazione e della variabilità dei caratteri, *Atti del R. Istituto Veneto di SS. LL. AA*, 73, 1203–1248.

GINI, C. (1921) Measurement of inequality and incomes, *The Economic Journal*, 31, 124–126.

GINI, C. and GALVANI, L. (1929) Di un'applicazione del metodo rappresentativo all' ultimo censimento Italiano della popolazione (1 dicembre 1921), *Annali di Statistica, Serie VI*, 4, 1–107.

GLASSER, G. (1962) Variance formulas for the mean difference and coefficient of concentration, *Journal of the American Statistical Association*, 57, 648–654.

HÁJEK, J. (1959) Optimum strategy and other problems in probability sampling, *Casopis pro Pestování Matematiky*, 84, 387–423.

HÁJEK, J. (1981) *Sampling from a Finite Population*, Marcel Dekker, New York.

HAMPEL, F. R., RONCHETTI, E., ROUSSEUW, P. J. and STAHEL, W. (1985) *Robust Statistics: The Approach Based on the Influence Function*, Wiley, New-York.

HANSEN, M H. (1987) Some history and reminiscences on survey sampling, *Statistical Science*, 2, 180–190.

HANSEN, M. H., DALENIUS, T. D., and TEPPING, B. J. (1985) *The development of sample survey in finite population*, In: A Celebration of Statistics (Eds. A. Atkinson and S. Fienberg), The ISI Centenary Volume, Springer, pp. 327-353.

HEDAYAT, A. S. and MAJUMDAR, D. (1995) Generating desirable sampling plans by the technique of trade-off in experimental design, *Journal of Statistical Planning and Inference*, 44, 237–247.

HOEFFDING, W. (1948) A class of statistics with asymptotically normal distribution, *Annals of Mathematical Statistics*, 19, 293–325.

HULLIGER, B. and MUNNICH, R. (2006) *Variance estimation for complex surveys in the presence of outliers*, In: Proceedings of the Section on Survey Research Methods, pp. 3153–3161, American Statistical Association.

HYNDMAN, R. J. and FAN, Y. (1996) Sample quantiles in statistical packages, *American Statistician*, 50, 361–365.

JENSEN, A. (1926) Report on the representative method in statistics, *Bulletin of the International Statistical Institute*, 22, 359–380.

KIAER, A. (1896) Observations et expériences concernant des dénombrements représentatifs, *Bulletin de l'Institut International de Statistique*, 9, 176–183.

KIAER, A. (1899) Sur les méthodes représentatives ou typologiques appliquées à la statistique, *Bulletin de l'Institut International de Statistique*, 11, 180–185.

KIAER, A. (1903) Sur les méthodes représentatives ou typologiques appliquées à la statistique, *Bulletin de l'Institut International de Statistique*, 31, 66–78.

KIAER, A. (1905) Discours sans intitulée sur la méethode repréesentative, *Bulletin de l'Institut International de Statistique*, 14, 119–134.

KOVACEVIC, M. S. and BINDER, D. A. (1997) Variance estimation for measures of income inequality and polarization - the estimating equations approach, *Journal of Official Statistics*, 13, 41–58.

KRUSKAL, W. and MOSTELLER, F. (1979A) Representative sampling, I: Nonscientific literature, *International Statistical Review*, 47, 13–24.

KRUSKAL, W. and MOSTELLER, F. (1979B) Representative sampling, II: Scientific literature, excluding statistics, *International Statistical Review*, 47, 111–127.

KRUSKAL, W. and MOSTELLER, F. (1979C) Representative sampling, III: The current statistical literature, *International Statistical Review*, 47, 245–265.

KRUSKAL, W. and MOSTELLER, F. (1980) Representative sampling, IV: The history of the concept in statistics, *International Statistical Review*, 48, 169–195.

LANGEL, M. and TILLÉ, Y. (2011) Statistical inference for the quintile share ratio, *Journal of Statistical Planning and Inference*, 141, 2976–2985.

LESAGE, E. (2008) *Contraintes d'équilibrage non linéaires*, In: Méthodes d'enquêtes : applications aux enquêtes longitudinales, à la santé et aux enquêtes électorales, pp. 285-289, (Eds. R. Guilbert, D. Haziza, D., A. Ruiz-Gazen, Y. Tillé), Dunod, Paris.

LORENZ, M. O. (1905) Methods of measuring the concentration of wealth, *Publications of the American Statistical Association*, 9, 209–219.

MONTI, A. C. (1991) The study of the Gini concentration ratio by means of the influence function, *Statistica*, 51, 561–577.

NEDYALKOVA, D. AND TILLÉ, Y. (2009) Optimal sampling and estimation strategies under linear model, *Biometrika*, 95, 521–537.

NEYMAN, J. (1934) On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, 97, 558–606.

NEYMAN, J. (1952) *Lectures and Conferences on Mathematical Statistics and Probability*, Graduate School; U. S. Department of Agriculture, Washington.

NYGARD, F. and SANDSTROM, A. (1985) The estimation of the Gini and the entropy inequality parameters in finite populations, *Journal of Official Statistics*, 1, 399–412.

OGWANG, T. (2000) A convenient method of computing the Gini index and its standard error, *Oxford Bulletin of Economics and Statistics*, 62, 123–129.

OSIER, G. (2006) *Variance estimation: the linearization approach applied by Eurostat to the 2004 SILC operation*, Technical report, Eurostat and Statistics Finland Methodological Workshop on EU-SILC, Helsinki, 7-8 November 2006.

OSIER, G. (2009) Variance estimation for complex indicators of poverty and inequality using linearization techniques, *Survey Research Methods*, 3,167–195.

PIGOU, A. C. (1912) *Wealth and Welfare*, McMillan, London.

ROYALL, R. M. and HERSON, J. (1973) Robust estimation in finite populations I, *Journal of the American Statistical Association*, 68, 880–889.

SANDSTROM, A., WRETMAN, J. H., and WALDEN, B. (1985) Variance estimators of the Gini coefficient: Simple random sampling, *Metron*, 43, 41–70.

SANDSTROM, A., WRETMAN, J. H., and WALDEN, B. (1988) Variance estimators of the Gini coefficient: Probability sampling, *Journal of Business and Economic Statistics*, 6, 113–120.

SHORROCKS, A. F. (1980) The class of additive decomposable inequality measures, *Economica*, 48, 613–625.

SHORROCKS, A. F. (1984) Inequality decomposition by population subgroups, *Econometrica*, 52, 1369–1385.

THEIL, H. (1967) *Economics and Information Theory*, Rand McNally.

THEIL, H. (1969) The desired political entropy, *The American Political Science Review*, 63, 521–525.

TILLÉ, Y. and FAVRE, A.-C. (2004) Co-ordination, combination and extension of optimal balanced samples, *Biometrika*, 91, 913–927.

TILLÉ, Y. and FAVRE, A. C. (2005) Optimal allocation in balanced sampling, *Statistics and Probability Letters*, 74, 31–37.

WOODRUFF, R. S. (1971) A simple method for approximating the variance of a complicated estimate, *Journal of the American Statistical Association*, 66, 411–414.

XU, K. (2003) *How has the literature on Gini's index evolved in the past 80 years?*, Working papers archive, Dalhousie, Department of Economics.

YATES, F. (1960) *Sampling Methods for Censuses and Surveys*, Charles Griffin, London, third edition.