The Gini Concentration Index in Survival Analysis

L'indice di concentrazione di Gini nell'analisi della sopravvivenza

Marco Bonetti[1], Chiara Gigliarano[2], Pietro Muliere[1]

[1] Department of Decision Sciences, Bocconi University, Italy
e-mail: marco.bonetti@unibocconi.it

[2] Department of Economics, Università Politecnica delle Marche, Italy

Riassunto: L'indice di concentrazione di Gini può essere utilizzato per misurare la concentrazione dei tempi di sopravvivenza di un insieme di pazienti, e per confrontare le distribuzioni di tempi di sopravvivenza in studi clinici. Qui presentiamo alcuni recenti risultati relativi alla definizione di uno stimatore per una versione ristretta dell'indice di Gini a partire da dati censurati a destra, al suo comportamento asintotico ed al suo utilizzo in particolare, nell'ambito dei modelli di cura.

Keywords: cure rate model, Gini index, linear rank test, right censored survival data

1. Introduction

The Gini index is one of the most common statistical indices employed in social sciences for measuring concentration in the distribution of a positive random variable; it is mainly used in economics as a measure of income or wealth inequality among individuals or households; see, e.g., Gini (1912, 1914), Nygard and Sandröm (1981), Kakwani (1980). Recently, the Gini coefficient has been used to describe concentration in levels of mortality, or in length of life, among different socio-economic groups, and to evaluate inequality in health and in life expectancy (see, e.g., Hanada, 1983 and Shkolnikov, Andreev and Begun, 2003).

Several tests based on the sample Gini index have been proposed in literature for noncensored data. Rao and Goria (2004) proposed a goodness-of-fit test that is based on the Gini index defined on spacings and showed, by simulation, that such test has higher power than all the competitors considered for certain common alternatives. Gail and Gastwirth (1978) and Nikitin and Tchirina (1996) considered tests of exponentiality based on the Gini index, also revealing high power against a broad class of alternatives. Niewiadomska-Bugaj, Kowalczyk and Ouda (2006) showed that a two-sample nonparametric test based on Gini index and applied to rank spacings can be more powerful than other tests based on rank spacings. The Gini concentration index has been studied in its ability to detect orderings in distributions (see for example Nygard and Sandröm, 1981 and Muliere and Scarsini, 1989). This, together with the good performance of Gini-based tests in the case of no censoring, motivates our proposal in this work of a two-sample nonparametric test based on a restricted version of such index for right censored data.

We focus in particular on cure rate models, i.e. survival models for which a fraction of the population never experience the event of interest; consequently, the corresponding survival function does not approach zero, but has a positive plateau.

A number of nonparametric statistical tests have been proposed in the literature to test the difference in survival distribution functions between groups, under the assumption

of positive cured rate. A common family of tests is the class of weighted linear rank tests, which includes the log-rank test, the Wilcoxon test, the Gray and Tsiatis test (see, e.g., Harrington and Fleming, 1982 and Gray and Tsiatis, 1989). Testing for differences between survival distributions via a concentration measure may prove more powerful than these methods, for example when one is far from the proportional hazard structure, or even in such a situation if the new test does not belong to the family of linear rank tests.

In the rest of this article we introduce an estimator for a restricted version of the Gini index, describe its asymptotic distribution, and discuss some aspects that are relevant to cure rate models. From these results we construct a new test for the equality of survival distributions.

2. The Gini concentration index

Consider a nonnegative random variable X with cumulative distribution function F, $F(x) = F_X(x) = P(X \leq x)$, survival function S, $S(x) = S_X(x) = 1 - F_X(x)$, probability density function $f(x)$, finite expected value $\mu = \int_{\Re^+} (1 - F(x))\, dx$ and variance $Var(X)$. Here we will focus on X as being a survival time.

The coefficient of mean difference is defined as

$$\delta = \int_{\Re^+} \int_{\Re^+} |x_1 - x_2|\, dF(x_1) dF(x_2). \tag{1}$$

The Gini coefficient of concentration for the distribution function F is defined as

$$G = \frac{\delta}{2\mu} \tag{2}$$

(see Gini, 1912, 1914, or Kendall and Stuart, 1977). The formula in (2) shows that the Gini index is invariant under scale transformations and depends on relative but not on absolute frequencies; also, it is bounded between 0 and 1. The minimum value is reached when everyone has the same length of life, while the maximum concentration is obtained when one individual has the entire "amount of life" and the rest of the population dies immediately.

Several other equivalent ways to define the Gini index exist. An alternative expression for the Gini coefficient, that will be used throughout this paper, is given by

$$\begin{aligned} G &= \frac{1}{\mu} \int_{\Re^+} F(u)(1 - F(u)) du = 1 - \frac{1}{\mu} \int_{\Re^+} (1 - F(u))^2 du \\ &= 1 - \frac{\int_{\Re^+} S^2(u) du}{\int_{\Re^+} S(u) du} \end{aligned} \tag{3}$$

(see, e.g., Michetti and Dall'Aglio, 1957 and Hanada, 1983).

The inferential aspects of estimators of the Gini coefficient G have also been established when data are characterized by no censoring. For each of the possible equivalent definitions of the Gini coefficient for a population distribution F, a sample counterpart exists, which can be computed on a sample $X_1, \ldots, X_n$ drawn from F. In particular, in Hoeffding (1948) the index G in (2) is estimated by the sample Gini concentration index $\widehat{G}_H$, given by

$$\widehat{G}_H = \frac{d}{2\overline{x}},$$

where d is the sample coefficient of mean difference $d = \frac{1}{n(n-1)} \sum_{j \neq k} |x_j - x_k|$ and $\overline{x}$ is the average of the realization $x_1, \ldots, x_n$ of the sample.

The quantity d is the arithmetic average of the $n(n-1)$ absolute differences in individual lengths of life; it assumes the maximum possible value equal to $2\bar{x}$ and the minimum equal to 0. Note that d is asymptotically equivalent to (1) computed on the empirical cumulative distribution function $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x)$ instead of $F(x)$, that is to the arithmetic mean of the n^2 terms $|x_j - x_k|$, including also the case of $j = k$.

The asymptotic distribution of $\widehat{G}_H$ has been discussed in Hoeffding (1948) as an application of his general results on U-statistics.

The Gini index is a natural candidate as a test for differences in two distribution functions when data are not censored; see, for example, the two-sample nonparametric test based on Gini index proposed in Niewiadomska-Bugaj, Kowalczyk and Ouda (2006).

As we are interested in applying the Gini index to lifetime data, in which individuals have finite follow-up time, it is natural to consider the following version of G:

$$G_t = 1 - \frac{\int_0^t S^2(u)du}{\int_0^t S(u)du}, \tag{4}$$

which we call the *restricted* Gini coefficient to distinguish it from the unrestricted Gini index G in (3) whose integrals in the definition run from zero to infinity. The time t represents the longest follow-up time in the data.

Note that there exist cases in which the restricted Gini coefficient G_t coincides with the unrestricted Gini index G, such as when the event of interest occurs before time t for all individuals. However, if at the end of follow-up there are still individuals who have not yet experienced the event of interest, then the survival function is positive at t and G_t will differ from G. In addition, G_t may also allow for the detection of differences in survival distributions in situation where the scale-invariance property would make G useless: for example, both for the exponential and for the Weibull distribution, the Gini coefficient calculated over the entire nonnegative real line assumes a constant value, independently of the hazard rate of the distribution. On the other hand, the restricted index G_t allows for comparison between such survival distributions.

The difference between G_t and G becomes particularly relevant whenever the length of follow-up does not allow the estimation of $S(u)$ until it gets close to zero, i.e. when follow-up is short compared to the survival times. This phenomenon is of interest in particular for cure rate models, since in those cases a fraction of the patient population never experiences the event, so that its survival times will necessarily be censored. These considerations suggest the study of the restricted Gini statistic within cure rate models, which we will discuss in Section 4.

3. Estimation and testing from right-censored data

We now turn to the problem of estimating G_t from sample censored data to test the hypothesis of equality of two survival distributions. The sample $X_1, \ldots, X_n$ is observed only partially, in particular after random right censoring. In other words, we do not observe the simple random sample $X_1, \ldots, X_n$ drawn from the distribution F, but rather the right censored sample $(\widetilde{X}_i, D_i)$, $i = 1, \ldots, n$, where $\widetilde{X}_i = X_i \wedge U_i$ and $D_i = 1(\widetilde{X}_i = X_i)$ for some censoring times $U_1, \ldots, U_n$, independently distributed as

Q and independent from the survival times. Then $N(u) = \sum_{i=1}^{n} 1(\widetilde{X}_i \leq u, D_i = 1)$ is a univariate counting process with intensity process $\beta(u) = \alpha(u)Y(u)$ (i.e., it satisfies the multiplicative intensity model structure), where the function $\alpha(u)$ is the hazard rate function at time u, i.e. $\alpha(u) = f(u)/S(u)$, and $Y(u) = \sum_{i=1}^{n} 1(\widetilde{X}_i \geq u)$ counts the number of subjects at risk just prior to time u. Here we focus on the nonparametric estimator of the survival function $S(u)$.

A natural estimator for G in (3) from right censored data can be constructed by replacing the Kaplan-Meier estimator $\widehat{S}(u)$ of $S(u)$ (see Kaplan and Meier, 1958 and Andersen, Borgan, Gill and Keiding, 1993). As we do not consider the quantity G that extends over the entire positive real line, but rather the quantities G_t obtained with the integrals in the definition running from zero to a maximum finite value $t < \infty$, we define an estimator of the restricted Gini index for right censored data as

$$\widehat{G}_t = 1 - \frac{\int_0^t \widehat{S}^2(u)du}{\int_0^t \widehat{S}(u)du}.$$

Use of $\widehat{G}_t$ requires at first the knowledge of its distributional properties. Note that the techniques based on U-statistics used in Hoeffding (1948) for the uncensored case do not apply here. Under some regularity conditions, one can show that the restricted Gini statistic for right censored data $\widehat{G}_t$ has a normal asymptotic distribution.

Theorem 1 *Consider the nonnegative function $y(u) = (1-F(u))(1-Q(u-))$ such that $\alpha(u)/y(u)$ is integrable over $[0,t]$, with $t < \infty$. For $s \leq t$ let $\sigma^2(s) = \int_0^s \frac{\alpha(u)}{y(u)}du$ and $J(s) = 1(Y(s) > 0)$, with $J(s)/Y(s) = 0$ whenever $Y(s) = 0$. Further, assume the following:*

(i) For each $s \in [0,t]$, $n\int_0^s \frac{J(u)}{Y(u)}\alpha(u)du \xrightarrow{\text{p}} \sigma^2(s)$ as $n \to \infty$.

(ii) For all $\epsilon > 0$, $n\int_0^t \frac{J(s)}{Y(s)}\alpha(s)I\left(\left|\sqrt{n}\frac{J(s)}{Y(s)}\right| > \epsilon\right)ds \xrightarrow{\text{p}} 0$ as $n \to \infty$.

(iii) $\sqrt{n}\int_0^t (1-J(s))\alpha(s)ds \xrightarrow{\text{p}} 0$ as $n \to \infty$.

Then, as $n \to \infty$,

$$\sqrt{n}\left(\widehat{G}_t - G_t\right) \xrightarrow{\text{d}} \mathcal{N}(0, \tau_t),$$

where

$$\tau_t = \int_0^t \left[4\frac{[\overline{\nu}_t(v)]^2}{W_t^2} + \frac{[\overline{\mu}_t(v)]^2 V_t^2}{W_t^4} - 4\frac{\overline{\mu}_t(v)\overline{\nu}_t(v)V_t}{W_t^3}\right] d\sigma^2(v), \tag{5}$$

with $\overline{\mu}_t(v) = \mu_t - \mu_v = \int_v^t S(u)du$, $\overline{\nu}_t(v) = \nu_t - \nu_v = \int_v^t S^2(u)du$, $W_t = \int_0^t S(u)du$ and $V_t = \int_0^t S^2(u)du$.

For a proof we refer to Bonetti, Gigliarano and Muliere (2008).

Note that when censoring gets stronger (say, from $Q(v) = Q_1(v)$ to $Q_2(v) \geq Q_1(v)\,\forall v$) the asymptotic variance τ_t increases, since $1-Q_2(v) \leq 1-Q_1(v)$ and therefore $d\sigma^2(v) = \alpha(v)/[(1-F(v))(1-Q(v-))]\,dv$ becomes larger in (5).

A plug-in estimator for the approximate variance of $\widehat{G}_t$ may be constructed by using uniformly consistent estimators for $\sigma^2(s)$, W_t, V_t, $\overline{\mu}_t$ and $\overline{\nu}_t$ into (5). While preparing this manuscript we have become aware that Tse (2006) has also proposed a nonparametric estimator for the Gini index when the data are randomly right censored and left truncated, deriving a central limit theorem for such Gini estimator. Differently from that work, which is based on the product-limit quantile process, our result provides a clearer expression for the asymptotic distribution of the Gini estimator, and an explicit formula for its asymptotic variance.

From the results of Theorem 1, we construct a test for comparing two survival functions, say S_1 and S_2, related, for example, to two different treatment groups. The Gini test that we propose is aimed to test for differences in two survival distributions from the point of view of concentration.

The Gini test statistic is

$$T := \frac{\left(\widehat{G}_{1,t} - \widehat{G}_{2,t}\right)^2}{\widehat{Var}(\widehat{G}_{1,t}) + \widehat{Var}(\widehat{G}_{2,t})}, \tag{6}$$

where $\widehat{G}_{j,t}$ is the estimator of the restricted Gini index for censored data referred to the treatment group j and $\widehat{Var}(\widehat{G}_{j,t})$ is the estimator of the approximate variance of $\widehat{G}_{j,t}$, for group j, $j = 1, 2$. For ease of notation we drop the indication of t from the name of the test statistic.

From Theorem 1 it follows that under the null hypothesis of equality of the two survival distributions, the statistic T has an approximate chi-squared distribution with 1 degree of freedom, while, under any alternative to the null hypothesis, T is distributed as an approximate noncentral chi-squared distribution with parameter of noncentrality $\eta = [(G_{1,t} - G_{2,t})^2]/(\tau_{1,t}/n_1 + \tau_{2,t}/n_2)$, where with obvious notation $\tau_{j,t}$ indicates the asymptotic variance of $\sqrt{n_j}\widehat{G}_{j,t}$ and n_j is the sample size of group j, $j = 1, 2$.

In applications, G_t can be estimated only up to the largest observed survival (or censoring) time. As these times (say $\tilde{t}_1$ and $\tilde{t}_2$) are typically different between the two treatment groups, the test statistic is constructed from $\widehat{G}_{1,\tilde{t}_1}$ and $\widehat{G}_{2,\tilde{t}_2}$.

As mentioned above, an interesting situation in which G_t differs from G is when the survival curve never approaches zero, i.e. under cure rate models. In the next section, we focus on that case.

4. The Gini concentration index in cure rate models

Cure rate models are survival models that typically study diseases with positive probability of being cured. They assume that the patient population can be split into two groups: the noncured patients, who experience the event of interest (relapse, death, etc.) before a given finite length of time, and the cured patients, who do not appear to be affected by the disease even after prolonged follow-up. Examples of diseases with positive cure rate are leukaemia, breast cancer, melanoma, head and neck cancer, non-Hodgkin's lymphoma, prostate cancer.

More specifically, cure rate models set a *cure time* t^* so that anyone who experiences the event of interest before t^* is considered a noncured patient, and anyone who survives after t^* is a cured patient. The fraction of cured patients is indicated with θ, $\theta \in [0, 1]$.

The portion $(1-\theta)$ of noncured patients experience the event of interest according to the conditional distribution function $F^*(u) = 1 - S^*(u)$, while the failure time of the cured patients is degenerate at infinity, i.e. their (improper) distribution function is constantly equal to 0. Therefore, the survival function of the entire patient population is given by

$$S(u) = \theta + (1-\theta)S^*(u).$$

Common choices for conditional survival function $S^*(u)$ are the exponential and the Weibull distributions. We call such models *traditional* cure rate models. Cure rate models have been studied extensively in the statistical literature; see, e.g., Gray and Tsiatis (1989), Ewell and Ibrahim (1997), Halpern and Brown (1987), Laska and Meisner (1992) and Sposto, Sather and Baker (1992).

Note that the unrestricted Gini coefficient G is not defined for distribution with infinite mean; however, one may consider a limit of the restricted Gini coefficient G_t, which is always defined. In particular, one can easily show that under a traditional cure rate model the restricted Gini index G_t in (4) converges to the proportion of noncured patients as t increases. This has a sample counterpart: as the follow-up time increases, the Gini test becomes equivalent to the proportions test, which is based on the difference of Kaplan-Meier estimates at time t. (See Bonetti, Gigliarano, and Muliere, 2008)

Several tests for the equality of two lifetime distributions within a cure rate model have been proposed in literature. Gray and Tsiatis (1989) considered a family of local alternatives based on differences in cure rates only. They compared distributions of failure time that are characterized by same conditional distributions for the noncured patients in both groups (i.e. $S_1^*(u) = S_2^*(u), \forall\, u$). They proved the optimality of their test within the class of weighted linear rank test of the type in Tarone and Ware (1977) and Gill (1980), with weight given by the left continuous version of the inverse of the Kaplan-Meier survival estimator. Ewell and Ibrahim (1997) discussed a more general class of local alternatives, according to which both the cure rates and the conditional distributions of the noncured patients differ between the two treatment groups. They derived the large sample distribution of the class of weighted log rank tests under such family of local alternatives and evaluated by simulations the power of the tests under the hypothesis of exponential conditional distributions F_1^* and F_2^*.

Due to lack of space we do not show here an illustration of the use of the test statistic in (6) to a clinical dataset, nor the results from an extensive simulation study within the setting of traditional cure rate models and generalized cure rate models (in which the distribution of the survival times of the cured patients is not degenerate at infinity). These will appear elsewhere.

However, the results from such simulation study suggest that the Gini index may be useful in some situations, and that it should be considered together with existing tests (in particular, the Log-rank, Wilcoxon, and Gray-Tsiatis tests).

5. Conclusions

We have summarized the recent results available on a restricted version of the Gini concentration coefficient for evaluating the degree of inequality in the distribution of the time to an event of interest within a given population. We described an estimator of the restricted index in presence of right censored data and its asymptotic distribution, and

proposed a consistent estimator of the corresponding asymptotic variance. We used these results to construct a new test for differences in survival distributions, with particular attention to cure rate models. We note that the test adjusts automatically to the censoring pattern in the data: in the case of no censoring it becomes equivalent to the proportions test and as such it puts all the importance on the cure rate; in case of censoring, it gives weight to differences in the conditional distributions of the noncured patients.

Clearly, rejection of the null hypothesis of equality of the survival distributions of two patient groups in a clinical study from the point of view of their concentration may suggest that subgroups of patients exist for whom treatment has a strong positive (or detrimental) effect, thus providing justification for further subgroups analyses.

Extensions of the use of the restricted Gini coefficient for censored data can be envisioned, for example via a polarization or inequality measure. In particular, polarization analysis detects for the presence in a given population of groups, which are similar inside and well separated from each other (see, e.g., Gigliarano, 2007). For a fixed number of groups an optimization method can be used to produce an optimal partition of the $\Re^+$ axis, so that the internal cohesion (as measured by the Gini index within the groups) is maximized (see Esteban, Gradín and Ray, 2007, and Aghevli and Mehran, 1981). Similarly, when working with right censored survival data one may look for an optimal partition of an interval $[0, t]$ such that the value of the restricted Gini coefficient within the partition is minimized. This would create a nonarbitrary partition into time intervals having homogeneous survival.

Further directions for research may also involve the analysis of dependence of survival on covariates, within the context of generalized Lorenz curves (see for example Muliere and Petrone, 1992).

References

Aghevli B.B., Mehran F. (1981) Optimal grouping of income distribution data, *Journal of the American Statistical Association*, 76, 22–26.

Andersen P.K., Borgan O., Gill R.D., Keiding N. (1993) *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

Bonetti M., Gigliarano C., Muliere P. (2008) The Gini concentration test for survival data, *in preparation*.

Esteban J., Gradín C., Ray D. (2007) An extension of a measure of polarization, with an application to the income distribution of five OECD countries, *Journal of Economic Inequality*, 5, 1–19.

Ewell M., Ibrahim J.G. (1997) The large sample distribution of the weighted log rank statistic under general local alternatives, *Lifetime Data Analysis*, 3, 5–12.

Gail M.H., Gastwirth J.L. (1978) A scale-free goodness-of-fit test for the exponential distribution based on the Gini statistic, *Journal of the Royal Statistical Society, B*, 40, 350–357.

Gigliarano C. (2007) *Polarization Measurement in One and More Dimensions*, Ph.D. Dissertation, Bocconi University, Milano, Italy.

Gill R.D. (1980) *Censoring and Stochastic Integrals*, Mathematisch Centrum, Amsterdam.

Gini C. (1912) Variabilità e mutabilità. Contributo allo studio delle distribuzioni e relazioni statistiche, *Studi Economico-Giuridici dell'Università di Cagliari*, III.

Gini C. (1914) Sulla misura della concentrazione e della variabilità dei caratteri, *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, LXXIII(part 2), 1203–1248.

Gray R.J., Tsiatis A.A. (1989) A linear rank test for use when the main interest is in differences in cure rates, *Biometrics*, 45, 899–904.

Halpern J., Brown B.W. (1987) Cure rate models: power of the logrank and generalized Wilcoxon tests, *Statistics in Medicine*, 6, 483–489.

Hanada K. (1983) A formula of Gini's concentration ratio and its applications to life tables, *Journal of Japanese Statistical Society*, 19, 293–325.

Harrington D.P., Fleming T.R. (1982) A class of rank test procedures for censored survival data, *Biometrika*, 69(3), 553–566.

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

Kakwani N.C. (1980) *Income Inequality and Poverty: Methods of Estimation and Policy Applications*, Oxford University Press, Oxford.

Kaplan E.M., Meier P. (1958) Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, 53, 457–481.

Kendall M., Stuart A. (1977) *The Advanced Theory of Statistics*, Vol. I, Mac Millan Publishing, New York.

Laska E.M., Meisner M.J. (1992) Nonparametric estimation and testing in a cure rate model, *Biometrics*, 48, 1223–1234.

Michetti B., Dall´Aglio G. (1957) La differenza semplice media, *Statistica*, 7, 159–255.

Muliere P., Petrone S. (1992) Generalized Lorenz curve and monotone dependence orderings, *Metron*, L(3-4), 19–38.

Muliere P., Scarsini M. (1989) A note on stochastic dominance and inequality measures, *Journal of Economic Theory*, 49, 314–323.

Niewiadomska-Bugaj M., Kowalczyk T., Ouda H. (2006) A new test of association and other tests based on the Gini mean difference, *Metron*, LXIV(3), 399–409.

Nikitin Y., Tchirina A.V. (1996) Bahadur efficiency and local optimality of a test for the exponential distribution based on the Gini statistic, *Journal of the Italian Statistical Society*, 5, 163–175.

Nygard F., Sandröm A. (1981) *Measuring Income Inequality*, Almqvist and Wilsell International, Stockholm.

Rao J.S., Goria M.N. (2004) A test of goodness-of-fit based on Gini's index of spacings, *Statistics and Probability Letters*, 68, 177–187.

Shkolnikov V.M., Andreev E.E., Begun A.Z. (2003) Gini coefficient as a life table function: computation from discrete data, decomposition of differences and empirical examples, *Demographic Research*, 8, 305–358.

Sposto R., Sather H.N., Baker S.A. (1992) A comparison of tests of the difference in the proportion of patients who are cured, *Biometrics*, 48, 87–99.

Tarone R.E., Ware J. (1977) On distribution-free tests for equality of survival distributions, *Biometrika*, 64, 156–160.

Tse S.M. (2006) Lorenz curve for truncated and censored data, *Annals of the Institute of Statistical Mathematics*, 58, 675–686.