International Journal of Pure and Applied Mathematics

Volume 25 No. 1 2005, 39-48

THE INFERENCE OF GINI'S MEAN DIFFERENCE

Ping-Huang Huang[1], Tea-Yuan Hwang[2 §]

[1]Department of Statistics and Insurance
Aletheia University
32 Chen-Li Street, Tamsui, Taipei, 25103, TAIWAN, R.O.C.
e-mail: phhuang@email.au.edu.tw

[2]Institute of Statistics
National Tsing Hua University
Hsinchu, 30013, TAIWAN, R.O.C.
e-mail: hwang@stat.nthu.edu.tw

Abstract: In this paper, the probability density function, the mean and the variance of Gini's mean difference are derived using the characterization of exponential distribution, and the procedures for its inference is presented and comparison with sample mean is done. More related results for references corresponding to gamma and Weibull distributions are studied and presented.

AMS Subject Classification: 26A33
Key Words: Gini's mean difference, Gini's statistics, Gini's concentration ratio, maximum likelihood estimator, gamma distribution, Weibull distribution, exponential distribution

1. Introduction

Let $X_1, X_2, \ldots, X_n$ be a set of samples drawn from F with sample mean $\overline{X}_n$. Then the statistic

$$G_n = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} |X_i - X_j|$$

is called *the Gini's mean difference* and $G_n^* = G_n/2\overline{X}_n$ is called *the Gini statis-*

Received: August 28, 2005

§Correspondence author

tic or *the Gini concentration ratio*. F are usually considered as exponential, Weibull and gamma distributions for lifetime distribution models. Actually it is relevant and widely used in different areas of statistics such as biochemistry and economics; 385 bibliographics references in Giorgi [6].

The distribution of G_n^* under exponentiality were first found by Girone [5]. Gail and Gastwirth [4] proved the asymptotic normality of G_n^* under the null hypothesis and alternatives. They also calculated the Pitman efficiency for gamma and Weibull families of alternatives that turned out to be fairly high (0.694 and 0.876 correspondingly). The sample scale-free Gini's statistics is known to be a powerful test of exponentiality against a broad class of alternatives. This test was considered by Gail and Gastwirth [4], see also Lawless [9], pp. 446-448.

Until now, the distribution of G_n even under exponentiality is still unknown. In this article, the characterization of gamma distribution or the independence of G_n^* and $\overline{X}_n$ which has been found by Hwang and Hu [8], Corolary 4.3, is used to derive its probability density function, the mean and the variance of G_n in Section 2. Its inference procedures under exponential distribution is shown and the comparison with $\overline{X}_n$ is done for small samples in Section 3. Some related results which is used in its inference under gamma and Weibull distributions are presented in Section 4.

2. The Distribution of Gini Mean Difference

Girone [5] first proved that G_n^* is distributed as $\overline{Y} = \frac{1}{n-1}\sum_{i=1}^{n-1} Y_i$, where $Y_1, Y_2, \ldots, Y_{n-1}$ are independent and uniformly distributed on (0, 1). Since $X_i = -\log Y_i$ has exponential population independently, and Bates [1] has obtained the probability density function (pdf) of $\overline{Y}$, thus the pdf of G_n^* is as the following:

$$f_{G_n^*}(g^*) = \frac{(n-1)^{n-1}}{(n-2)!} \sum_{j=0}^{[(n-1)g^*]} (-1)^j \binom{n-1}{j} \left(g^* - \frac{j}{n-1}\right)^{n-2}, \qquad (2.1)$$

where $[\cdot]$ is Gaussian integer and $0 \le g^* \le 1$.

Since Hwang and Hu [8], Corollary 4.3, proved that the independence of G_n^* and $\overline{X}_n$ is equivalent to exponential distribution with scale parameter λ, then G_n has the following probability density function:

$$f_{G_n}(g) = \int_0^\infty \frac{(\lambda n)^n}{2\Gamma(n)} \overline{x}^{(n-2)} e^{-n\lambda\overline{x}} \cdot \frac{(n-1)^{n-1}}{(n-2)!} \sum_{j=0}^{[(n-1)g/2\overline{x}]} (-1)^j \binom{n-1}{j} \times \left(\frac{g}{2\overline{x}} - \frac{j}{n-1}\right)^{n-2} d\overline{x} = \frac{n\lambda}{2\Gamma(n-1)} \Big[\sum_{j=1}^{n-2} c_{n,j} \exp\left(-\frac{n(n-1)\lambda g}{2j}\right) - \left(\sum_{j=1}^{n-2} c_{n,j}\right) \exp\left(-\frac{n\lambda g}{2}\right) \Big], \quad (2.2)$$

where $g > 0$, $c_{n,j} = (-1)^{n+j-1} c_j^{n-1} j^{n-2}$. The last equation is derived by mathematical induction.

In this paper, the mean and the variance of G_n under exponential distribution are derived using (2.2) as follows:

$$E(G_n) = \frac{2}{\lambda(n-1)n!} \sum_{j=1}^{n-2} c_{n,j}(j^2 - (n-1)^2), \quad (2.3)$$

where $c_{n,j}$ is defined as in (2.2) and

$$\begin{aligned} \mathrm{Var}(G_n) &= E(G_n^2) - E^2(G_n) \\ &= \frac{1}{\lambda^2} \left[\frac{8(-1)^n}{n^2(n-1)^2} \sum_{j=1}^{n-2} (-1)^j \frac{j^n}{j!(n-2-j)!} + \frac{4}{n} - 1 \right]. \end{aligned} \quad (2.4)$$

The $\mathrm{Var}(G_n) > 0$ for $n \geq 3$ would be verified as follows: when n is odd, there are $(n-2)$ or odd terms in the summation, and two terms are combined pairwise from second term, thus the sum of each pair is negative and the sum of the summation is also negative since the first term is negative and $\mathrm{Var}(G_n) > 0$ since $(-1)^n$, the first term and $(4/n - 1)$ are all negative. When n is even, similar procedure can be applied to obtain the same result. Thus $\mathrm{Var}(G_n) > 0$ for $n \geq 3$ is established.

The mean and the variance of Gini's mean difference G_n under exponential distribution have been derived by Nair [11], Lomnicki [10] and David [2] as follows:

$$E(G_n) = \frac{1}{\lambda} \text{ and } \mathrm{Var}(G_n) = \frac{2(2n-1)}{3n(n-1)\lambda^2}. \quad (2.5)$$

Combing (2.3) and (2.4) with (2.5) under exponential distribution, the following identity can be drawn:

$$\sum_{j=1}^{n-2} c_{n,j}(j^2 - (n-1)^2) = \frac{n!(n-1)}{2} \quad (2.6)$$

where $c_{n,j}$ is defined in (2.2) and

$$\frac{24(-1)^n}{n(n-1)!}\sum_{j=1}^{n-2}(-1)^j c_j^{n-2} j^n = 3n^2 - 11n + 10 \tag{2.7}$$

3. The Inference Procedures for Exponential Distribution

Hoffding [7] has proved that G_n is asymptotically normal under any population, so both sample mean $\overline{X}_n$ and G_n are asymptotically normal under exponential distribution. Could we conclude that $\overline{X}_n$ is better than G_n by $\mathrm{Var}(\overline{X}_n) < \mathrm{Var}(G_n)$ for all n? According to the distribution of $\overline{X}_n$ is more skew than exponential distribution; more skew as n increases, therefore $\overline{X}_n$ is not necessarily more reliable than G_n for small samples. For proving this fact, we calculate the 95% confidence intervals (C.I.) of G_n and $\overline{X}_n$ respectively for various σ and n by using simulation study and then compare their lengths. The percentages of C.I. of G_n shorter than that of $\overline{X}_n$ are presented in the following Table 1.

From Table 1, conclusion can be more accurately drawn that Gini's mean difference G_n will be more reliable than sample mean $\overline{X}_n$ for σ is greater than 1.33, 1.17, 1.14 and 1.12 or λ less than 0.75, 0.85, 0, 88 and 0.89 when $n = 5$, 10, 15 and 20 respectively. From above conclusion, the fact holds true for σ and λ tending to 1 when n increases to infinite.

Confidence intervals and tests for λ or σ of exponential distribution are easily obtained using the pivotal quantity λG_n. For example, to obtain an equitail, two-sided $1-\alpha$ confidence interval for λ (constant hazard rate), we take

$$\Pr(g_{\alpha/2,n} \le \lambda G_n \le g_{1-\alpha/2,n}) = 1 - \alpha\,, \tag{3.8}$$

where $g_{p,n}$ satisfies $\Pr(\lambda G_n \le g_{p,n}) = p$. Then

$$\frac{g_{\alpha/2,n}}{G_n} \le \lambda \le \frac{g_{1-\alpha/2,n}}{G_n} \tag{3.9}$$

is the $1-\alpha$ confidence interval for λ. The p-th quantile $g_{p,n}$ of the distribution of λG_n can be found for $3 \le n \le 20$ from the following Table 2.

Note that $g_{p,n}$ is also the p-th quantile of G_n with $\lambda = 1$, and obviously the p-th quantile $g_{p,n}(\lambda)$ of G_n with $\lambda \ne 1$ is $g_{p,n}/\lambda$.

Similarly, G_n can be used as a test statistic for hazard rate λ for exponential distribution. Formally, suppose we wish to test $H_0 : \lambda = \lambda_0$ verses $H_1 : \lambda >$

$\sigma\backslash n$	5	10	15	20
0.30	0.1	0	0	0
0.40	0.7	0	0	0
0.50	0.9	0	0	0
0.60	1.8	0.2	0	0
0.70	3.6	0.5	0.1	0.1
0.80	7.1	2.6	0.7	0
0.90	11.6	7.7	3.1	1.4
1.00	19	16.9	12.8	10.6
1.12	29.7	36.3	42.6	54.2
1.14	30.6	44.7	52.8	61.6
1.15	30.8	46.2	54.9	67.2
1.17	34.8	49.7	61.8	72.4
1.30	45	78.4	94.1	99.4
1.33	50.6	84.4	97.0	99.8
1.60	80.2	99.9	100.0	100.0
1.68	85.7	100.0	100.0	100.0
2.20	100.0	100.0	100.0	100.0

Table 1: The percentage of 95% C. I. of G_n shorter than that of $\overline{X}_n$

λ_0. In this case large values of G_n provide evidence against H_0; the observed significance level (p value) for H_0 verses H_1 is $p = \Pr(G_n \geq g_n : \lambda = \lambda_0) = \Pr(\lambda_0 G_n \geq \lambda_0 g_n)$, where g_n is the observed value of G_n. Thus H_0 is rejected when $p < \alpha$, where α is the significance level.

Since the Gini's mean difference is introduced as a measure of variation, thus G_n can be used as an estimator of the standard deviation $\sigma(= \frac{1}{\lambda})$ of exponential population and the confidence interval of σ can be found from (3.1) as following:

$$\frac{G_n}{g_{1-\alpha/2,n}} \leq \sigma \leq \frac{G_n}{g_{\alpha/2,n}}.$$

Similarly, G_n can be used as a test statistic for standard deviation σ mentioned as before.

The inference for some other characteristics of the distribution can be similarly proceed, since these are simple function of λ. In particular, confidence intervals and tests are readily obtained for (1) the survivor (i.e. reliability) function at time t_0, given by $S(t_0) = \exp(-\lambda t_0)$ and (2) the p-th quantile of the distribution, given by $g_{p,n}(\lambda) = \lambda F_{G_n}^{-1}(p)$. By (3.1), the interval (S_L, S_U)

Quantile							
n	0.01	0.025	0.05	n	0.95	0.975	0.99
3	0.07024	0.11474	0.16873	3	2.45077	2.91717	3.53060
4	0.13365	0.19056	0.25321	4	2.19796	2.56093	3.03278
5	0.18673	0.24909	0.31469	5	2.04128	2.34379	2.73356
6	0.23080	0.29573	0.36220	6	1.93198	2.19427	2.52984
7	0.26802	0.33412	0.40053	7	1.85025	2.08360	2.38090
8	0.30001	0.36650	0.43238	8	1.78619	1.99750	2.26510
9	0.32789	0.39433	0.45942	9	1.73422	1.92826	2.17293
10	0.35251	0.41860	0.48279	10	1.69107	1.87104	2.09719
11	0.37445	0.44003	0.50325	11	1.65451	1.82279	2.03360
12	0.39418	0.45914	0.52138	12	1.62301	1.78140	1.97929
13	0.41205	0.47633	0.53759	13	1.59553	1.74542	1.93225
14	0.42835	0.49191	0.55220	14	1.57127	1.71378	1.89101
15	0.44328	0.50611	0.56546	15	1.54967	1.68569	1.85450
16	0.45704	0.51913	0.57757	16	1.53030	1.66057	1.82190
17	0.46978	0.53113	0.58869	17	1.51274	1.63784	1.79258
18	0.48161	0.54224	0.59895	18	1.49679	1.61724	1.76601
19	0.49265	0.55256	0.60845	19	1.48219	1.59844	1.74181
20	0.50298	0.56219	0.61730	20	1.46878	1.58119	1.71967

Table 2: Quantiles $g_{p,n}$ of the test statistics λG_n

is an $(1-\alpha)$ confidence interval for $S(t_0)$ where $S_L = \exp(-t_0 g_{\alpha/2,n}/G_n)$ and $S_U = \exp(-t_0 g_{1-\alpha/2,n}/G_n)$, and

$$(g_{\alpha/2,n} F^{-1}_{G_n}(p)/G_n, g_{1-\alpha/2,n} F^{-1}_{G_n}(p)/G_n)$$

is an $(1-\alpha)$ confidence interval for $g_{p,n}(\lambda)$. The procedures for testing $S(t_0)$ and $g_{p,n}(\lambda)$ are similar as mentioned above.

4. Some Related Results

By Hoeffding [7] and Central Limit Theorem, it is necessary to find $E(G_n)$ and $\mathrm{Var}(G_n)$ under any distribution. In this section, the G_n used to do inference under gamma and Weibull distributions will be discussed.

The mean and the variance of Gini's mean difference have been derived by Nair [11], Lomnicki [10] and David [2] under any distribution with different approaches as follows:

Distribution	P.D.F.	$E(G_n)$	$\mathrm{Var}(G_n)$
Weibull	$\lambda\beta(\lambda x)^{\beta-1}e^{-(\lambda x)^{\beta}}$, x, λ, $\beta > 0$	$(2-2^{1-\frac{1}{\beta}})\frac{\Gamma(1+\frac{1}{\beta})}{\lambda}$	$(*)$
Gamma	$\frac{\lambda(\lambda x)^{k-1}e^{-\lambda x}}{\Gamma(k)}$, x, λ, $k > 0$	$\frac{2k}{\lambda} - \frac{k}{\lambda 2^{k-1}}\sum_{i=0}^{k-1}\frac{c_i^{k+i}}{i!}$	$(**)$

Table 3: The expectation and the variance of G_n for Weibull and gamma distributions

$$E(G_n) = 4\int xF(x)dF(x) - 2\int xdF(x)$$
$$= 2\int xdF(x) - 4\int x(1-F(x))dF(x)$$

and

$$\mathrm{Var}(G_n) = \frac{1}{n(n-1)}\{4(n-1)\sigma^2 + 16(n-2)I - 2(2n-3)E^2(G_n)\},$$

where

$$I = \int\{[xF(x) - Z(x)]^2 + (\mu - x)[xF(x) - Z(x)]\}f(x)dx,$$

with μ the mean and σ^2 variance of X, and $Z(x) = \int_{-\infty}^{x} tf(t)dt$.

Since Gini's mean difference can be rewritten as $G_n = \frac{2}{n(n-1)}\sum_{i=1}^{n}(2i - n - 1)X_{(i)}$ (cf. David [3], p. 167), and if rewriting $(2i-n-1)$ as $(i-1)-(n-i)$, we have the easier procedure comparing with previous three authors to calculate $E(G_n)$ and $\mathrm{Var}(G_n)$. Note that $\mathrm{Var}(G_n)$ depends on n but $E(G_n)$ does not. In this paper, $E(G_n)$ and $\mathrm{Var}(G_n)$ are derived corresponding to some well-known distributions and presented in the following Table 3; the $(*)$ and $(**)$ denote the variances corresponding to Weibull and gamma distributions respectively. Note that both $(*)$ and $(**)$ do reduce to the variance of exponential distribution for $k = 1$.

$(*)$ $\mathrm{Var}(G_n) = \frac{1}{n(n-1)}\{4(n-1)\sigma^2 + 16(n-2)I_W - 2(2n-3)E^2(G_n)\}$, where $E(G_n)$ is presented in Table 3 and

$$I_W = u^2 + \frac{\Gamma\left(\frac{2}{\beta}+1\right)}{\lambda^2}\left(3^{-\frac{2}{\beta}-1} - 2^{-\frac{2}{\beta}-1}\right) - \frac{u}{\lambda}\frac{\Gamma\left(\frac{1}{\beta}+1\right)}{2^{\frac{2}{\beta}+1}}$$

$$+ W_{0,2,1} + W_{1,1,1} - 2W_{1,1,2} + W_{0,1,1}$$

with

$$W_{i,j,k} = \int_0^\infty \beta\lambda^\beta x^{\beta-1+i} Z^j(x) e^{-k(\lambda x)^\beta} dx \quad \text{and} \quad Z(x) = \frac{1}{\lambda}\int_0^{(\lambda x)^\beta} y^{\frac{1}{\beta}} e^{-y} dy.$$

$$(**)\ \mathrm{Var}(G_n) = \frac{1}{n(n-1)}\{4(n-1)\frac{k}{\lambda}^2 + 16(n-2)I_G - 2(2n-3)E^2(G_n)\},$$

where $E(G_n)$ is presented in Table 3 and

$$I_G = \frac{k^2}{\lambda^2}\left(\frac{1}{3^k} - \frac{1}{2^{k+1}}\right) + \frac{1}{\lambda^2\Gamma(k)}B\,,$$

with

$$B = \Big\{\sum_{i=1}^{k} C_i^2 \frac{\Gamma(k+2i)}{3^{k+2i}} + \sum\sum_{i\neq j} C_i C_j \frac{\Gamma(k+i+j)}{2^{k+i+j-1}}, +2k\sum_{i=1}^{k} C_i \frac{\Gamma(k+i)}{3^{k+i}}\Big\}$$

and $C_i = \frac{1}{(i-1)!}\left(\frac{k}{i} - 1\right)$.

The closed form of the moments of G_n^* is still unknown. Although it is easy to use its probability density function directly to derive them for particular n, it is very complicated for large n. However, the closed form of the mean and variance of G_n^* can be derived easily by using the characterization of gamma or generalized gamma distribution obtained by Hwang and Hu [8], Corollary 4.3, as follows: Since the independence of G_n^* and $2\overline{X}_n$ gives $E(G_n) = 2E(G_n^*)E(\overline{X}_n)$, then we have

$$E(G_n^*) = \frac{1}{\lambda}\Big/\frac{2}{\lambda} = \frac{1}{2}$$

under exponential distribution, $E(G_n^*) = 1 - 2^{-2+\frac{1}{\beta}}$ under Weibull distribution and $E(G_n^*) = 1 - \sum_{i=0}^{k-1} \frac{c_i^{k+1}}{i!2^k}$ under gamma distribution.

Furthermore $G_n^{*^r}$ and $\overline{X}_n^r$ are also independent, then $E(G_n^r) = 2^r E(G_n^{*^r})$ $E(\overline{X}_n^r)$ follows by the characterization mentioned above. Since $E(\overline{X}_n^r)$ is known, thus $E(G_n^r)$ or $E(G_n^{*^r})$ can be found when $E(G_n^{*^r})$or $E(G_n^r)$ is known for $r > 0$, and the r-th central moment of G_n or G_n^* can be found.

For example, Gail and Gastwirth [4] proved that $E(G_n^*) = \frac{1}{2}$ and $\mathrm{Var}(G_n^*)$ $= \frac{1}{12(n-1)}$ under exponential distribution, then $E(G_n^{*^2}) = (3n-2)/12(n-1)$ and

$$E(G_n^2) = 4 \cdot \left(\frac{3n-2}{12(n-1)}\right)\left(\frac{n+1}{n\lambda^2}\right) = \frac{(3n-2)(n+1)}{3n(n-1)\lambda^2}. \tag{4.1}$$

Distribution	P.D.F.	$E(G_n^*)$	$\mathrm{Var}(G_n^*)$
Weibull	$\lambda\beta(\lambda x)^{\beta-1}e^{-(\lambda x)^\beta}$, x, λ, $\beta > 0$	$1-2^{-\frac{1}{\beta}}$	$(*)$
Gamma	$\frac{\lambda(\lambda x)^{k-1}e^{-\lambda x}}{\Gamma(k)}$, x, λ, $k > 0$	$1-\frac{1}{2^k}\sum_{i=0}^{k-1}\frac{c_i^{k+1}}{i!}$	$(**)$

Table 4: The expectation and the variance of G_n^* for Weibull and gamma distributions

Thus

$$\mathrm{Var}(G_n) = E(G_n^2) - E^2(G_n) = \frac{2(2n-1)}{3n(n-1)\lambda^2} \tag{4.2}$$

under exponential distribution, which is as same as (2.5). Furthermore, $E(G_n^*)$ and $\mathrm{Var}(G_n^*)$ can be derived for both Weibull and gamma distributions by the similar procedure as exponential distribution. The results are presented as the following Table 4.

$$(*)\ \mathrm{Var}(G_n^*) = \frac{n\lambda^2[\mathrm{Var}(G_n) + E^2(G_n)]}{4[\Gamma(1+\frac{2}{\beta}) + (n-1)\Gamma^2(1+\frac{1}{\beta})]} - E(G_n^*)\,,$$

where $\mathrm{Var}(G_n)$ and $E(G_n)$ are presented in Table 3, $E(G_n^*)$ in Table 4.

$$(**)\ \mathrm{Var}(G_n^*) = \frac{n\lambda^2[\mathrm{Var}(G_n) + E^2(G_n)]}{4[\Gamma(1+\frac{2}{\beta}) + (n-1)\Gamma^2(1+\frac{1}{\beta})]} - E(G_n^*)\,,$$

where $\mathrm{Var}(G_n)$ and $E(G_n)$ are presented in Table 3, $E(G_n^*)$ in Table 4

$$\mathrm{Var}(G_n^*) = \frac{n}{nk+1}[(n-1)k + 4(n-2)I_{G^*} + (6k-4nk-n+1)E^2(G_n^*)\}$$

where $E(G_n^*)$ is presented in Table 4 and

$$I_{G^*} = k\left(\frac{1}{3^k} - \frac{1}{2^{k+1}}\right) + \frac{k}{\Gamma(k)}B$$

with B defined in Table 3.

References

[1] G.E. Bates, Joint distributions of time intervals for the occurrence of successive accidents in a generalized P'olva scheme, *Ann. Math. Statist.*, **26** (1955), 705-720.

[2] H.A. David, Gini's mean difference rediscovered, *Biometrika*, **55** (1968), 573-575.

[3] H.A. David, *Order Statistics*, Wiley, New York (1981).

[4] M.H. Gail, J.L. Gastwirth, A scale-free goodness-of-fit test for the exponential distribution based on the Gini statistics, *J. Roy. Statist. Soc. Ser. B*, **40**, No. 3 (1978), 350-357.

[5] G. Girone, La distriluzione del rapporto di concentrazione per campioni casuali di variabili esponenziali, In: *Studi di Probabilita, Statistica e Ricerca Operation in Orone di Giuseppe Pompilj* (Ed. G. Dall'Aglio) Oderisi, Gubbio (1971), 320-326.

[6] G.M. Giorgi, Bibliographic portait of the Gini concentration ration, *Metron*, **XLVIII**, No. 1-4 (1990), 183-221.

[7] W. Hoeffding, A class of statistics with asymptotically normal distribution, *Ann. Math. Statist.*, **19** (1948), 293-325.

[8] T.Y. Hwang, C.Y. Hu, On some characterizatios of population distributions, *Taiwanese J. Math.*, **4** (2000), 427-437.

[9] J.F. Lawless, *Statistical Models and Methods for Lifetime Data*, Wiley, New York (1982).

[10] Z.A. Lomnicki, The standard error of Gini's mean difference, *Ann. Math. Statist.*, **23** (1952), 635-637.

[11] U.S. Nair, The standard error of Gini's mean difference, *Biometrika*, **28** (1936), 428-436.

[12] J. Neyman, Outline of a theory of statistical estimation based on the classical theory of Probability, *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.,* **236** (1937), 333.

[13] S.S. Wilks, Shortest average confidence intervals from large samples, *Ann Math. Statist.*, **9** (1938), 166.