BIOMATHEMATICS TRAINING PROGRAM

GINI'S MEAN DIFFERENCE AS A NONPARAMETRIC MEASURE OF SCALE

by

Dennis D. Boos

Department of Statistics
North Carolina State University

Institute of Statistics Mimeo Series # 1166

April, 1978

Gini's mean difference as a nonparametric measure of scale

By Dennis D. Boos

Department of Statistics, North Carolina State University, Raleigh

SUMMARY

We propose Gini's mean difference as a highly competitive alternative to the pth power deviations suggested by Bickel and Lehmann (1976). Comparisons are based on standardized asymptotic variances and Monte Carlo simulation.

Some key words: Gini's mean difference; Pth power deviation; Standardized asymptotic variance; Test for scale; Jackknife.

1. INTRODUCTION

Bickel and Lehmann (1976) proposed the pth power deviations $\tau_p = (E|X-\mu|^p)^{1/p}$ as a suitable family of scale measures for symmetric distributions. These measures satisfy certain "measures of scale" criteria and can be estimated with fairly high asymptotic efficiency over a wide range of distributions. Gini's mean difference $\Delta = E|X_1-X_2|$ also satisfies the scale measure criteria, and we seek to compare Δ to several selected members of the pth power family.

The estimator of Δ, also called Gini's mean difference, can be defined for a sample $X_1,\ldots,X_n$ as a U-statistic

$$\hat{\Delta} = \frac{2}{n(n-1)} \sum_{i<j} |X_i - X_j|$$

or as an L-statistic

$$\hat{\Delta} = \frac{2}{n(n-1)} \sum_{i=1}^{n} (2_i - n - 1) X_{in} ,$$

where $X_{1n} \leq X_{2n} \leq \ldots \leq X_{nn}$ is the ordered sample. David (1968) gives a brief history of $\hat{\Delta}$ and traces its origin back to von Andrae (1872). Perhaps $\hat{\Delta}$ is best known as a highly efficient estimator of $(2/\sqrt{\pi})\sigma$ in normal samples; see, e.g., Nair (1936), Downton (1966) and D'Agostino (1970). More recently $\hat{\Delta}$ has been used by D'Agostino (1971) in a test for normality and by Wainer and Thissen (1976) in the construction of a robust estimator of correlation.

In Section 2 we show that Δ satisfies the Bickel and Lehmann criteria for scale measures, and then for selected distributions we compare the standardized asymptotic variance of $\hat{\Delta}$ to that of $\hat{\tau}_2$ and $\hat{\tau}_1$, the sample standard deviation and mean deviation. In Section 3 we construct two sample tests for scale based on log $\hat{\Delta}$ and log $\hat{\tau}_p$ a la Miller (1968). Monte Carlo simulations are used to make small sample comparisons.

2. SCALE MEASURES

For X having distribution F, let the notation $\sigma(X)$ or $\sigma(F)$ refer to a measure evaluated at F such as the standard deviation, $\tau_2(X) = (E|X-\mu|^2)^{\frac{1}{2}}$. Let $X \overset{st}{\geq} Y$ mean that X is stochastically larger than Y, i.e., $P(X > x) \geq P(Y > x)\ \forall x$. The three basic criteria suggested by Bickel and Lehmann (1976) for a scale measure are (1) $\sigma(aX) = |a|\sigma(X)$ $a > 0$, (2) $\sigma(X+b) = \sigma(X)\ \forall\ b$ and (3) for symmetric distributions

$$|Y-\mu_Y| \overset{st}{\geq} |X-\mu_X| \Rightarrow \sigma(Y) \geq \sigma\ (X)\ , \qquad (2.1)$$

where μ_Y and μ_X are the centers of symmetry. Clearly $\Delta = E|X_1-X_2|$ satisfies (1) and (2). Under the assumption of symmetric unimodal densities g and f for independent Y_i, X_i, i = 1,2, Theorem 1 of Bickel and Lehmann (1976) yields

$$|Y-\mu_Y| \overset{st}{\geq} |X-\mu_X| \Rightarrow |Y_1-Y_2| \overset{st}{\geq} |X_1-X_2|\ .$$

Then if g and f have finite means, criterion (3) follows for Δ from the integration by parts identity

$$E(|Y_1-Y_2| - |X_1-X_2|) = \int_0^\infty \{F_{|X_1-X_2|}(x) - F_{|Y_1-Y_2|}(x)\}\, dx\ .$$

Note that the same proof applies to the family $\Delta_p = (E|X_1-X_2|^p)^{1/p}$; see Katti (1960) for Δ_p in discrete distributions. The Δ_p family does not involve a measure of location, except $\Delta_2 = \sqrt{2}\,\tau_2$ implicitly, and motivates the replacement of (2.1) by

$$|Y_1-Y_2| \overset{st}{\geq} |X_1-X_2| \Rightarrow \sigma(Y) \geq \sigma(X)$$

for general, possibly asymmetric distributions. The family of measures $\sigma_J(F) = \int_0^1 F^{-1}(t)J(t)dt$, where $J(t)$ is skew-symmetric about $\frac{1}{2}$, also contains Δ, $J(t) = 4(t-\frac{1}{2})$, and satisfies the three Bickel and Lehmann criteria for scale measures. This family has been extensively studied in particular parametric settings and allows for flexibility regarding trimming and censoring; see Johnson and Kotz (1970, Vol. 1, p. 66-72). However, we want to restrict our attention to Δ and consider a wide range of distributions.

In Table 1 we list the actual values of Δ and the ratio Δ/τ_2 for the following distributions: uniform $(0,b)$, normal $(0,\sigma^2)$, logistic - $F(x) = (1+e^{-x/\sigma})^{-1}$, Laplace - $f(x) = (2\sigma)^{-1}e^{-|x|/\sigma}$, exponential - $f(x) = \sigma^{-1}e^{-x/\sigma}I(0 \leq x)$.

Table 1. Values of Gini's mean difference Δ and its ratio to the standard deviation, Δ/τ_2.

	Uniform	Normal	Logistic	Laplace	Exponential
Δ	$\frac{b}{3}$	$\frac{2\sigma}{\sqrt{\pi}}$	2σ	$\frac{3\sigma}{2}$	σ
Δ/τ_2	1.155	1.128	1.103	1.061	1

The variance of $\hat{\Delta}$ is easily calculated to be

$$\mathrm{Var}(\hat{\Delta}) = \frac{1}{n}\left[\frac{4\sigma^2-4J+2\Delta^2}{n-1} + 4(J-\Delta^2)\right],$$

where σ^2 is the variance of X and $J = E|X_1-X_2||X_1-X_3|$. Small sample efficiencies have been tabulated for several distributions by Nair (1936) and Sarhan (1954, 1955). However, for space considerations and ease of computation we prefer to look at the <u>standardized asymptotic variance</u>, defined for estimators $\hat{\theta}$ which are asymptotically normal with mean θ and variance A^2/n by

$$sv(\hat{\theta}) = A^2/\theta^2 .$$

Not only is this measure invariant with respect to scaling factors, but $sv(\hat{\theta})/n$ is the asymptotic variance of $\log \hat{\theta}$. This latter fact provides a bridge to the test statistics found in Section 3. We define the <u>standardized asymptotic relative efficiency</u> of $\hat{\theta}_1$ to $\hat{\theta}_2$ by $\mathrm{eff}(\hat{\theta}_1,\hat{\theta}_2) = sv(\hat{\theta}_2)/sv(\hat{\theta}_1)$. In Table 2 we list the standardized asymptotic variances for the mean difference $\hat{\Delta}$, the sample standard deviation $\hat{\tau}_2$, and for the sample mean deviation from the mean $\hat{\tau}_1$. Also listed are the standardized asymptotic relative efficiencies of $\hat{\Delta}$ to $\hat{\tau}_2$ and to $\hat{\tau}_1$.

A truly remarkable fact is that the efficiencies of $\hat{\Delta}$ with respect to the maximum likelihood estimators for the normal, logistic, and Laplace families are .978, .985, and .964 respectively. Even in the highly skewed exponential the efficiency of $\hat{\Delta}$ with respect to the mean is .75. In Table 3 we list the efficiencies of $\hat{\Delta}$ with respect to the standard deviation $\hat{\tau}_2$ for various members of the Tukey normal gross error distributions, $F(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/\lambda)$, where Φ is the standard normal. This table is to be compared with Table 5.3 of Bickel and Lehmann (1976) which lists eff $(\hat{\tau}_1, \hat{\tau}_2)$ and eff $(\hat{\tau}_{1.5}, \hat{\tau}_2)$ for the same (ε, λ) combinations. For these distributions $\hat{\Delta}$ appears to perform better than $\hat{\tau}_2$ and $\hat{\tau}_{1.5}$, but not quite as well as $\hat{\tau}_1$.

Table 2. Standardized asymptotic variances and relative efficiencies for Gini's mean difference $\hat{\Delta}$, the standard deviation $\hat{\tau}_2$, and for the mean deviation $\hat{\tau}_1$.

	Uniform	Normal	Logistic	Laplace	Exponential
$sv(\hat{\Delta})$	.200	.511	.710	1.037	1.333
$sv(\hat{\tau}_2)$	.200	.500	.800	1.250	2.000
$sv(\hat{\tau}_1)$	.333	.571	.712	1.000	1.437
$eff(\hat{\Delta}, \hat{\tau}_2)$	1.000	.978	1.127	1.205	1.500
$eff(\hat{\Delta}, \hat{\tau}_1)$	1.667	1.117	1.003	.964	1.077

Table 3. Asymptotic relative efficiency of Gini's mean difference with respect to the standard deviation for Tukey normal gross error distributions, $F(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/\lambda)$.

$\varepsilon \backslash \lambda$	2	4	6
0	.978	.978	.978
.025	1.122	2.581	3.746
.075	1.222	1.975	1.957
.10	1.231	1.742	1.624
.20	1.190	1.262	1.108
.40	1.077	.991	.900
.50	1.039	.952	.884

Finally we note that Δ is not robust in the strict Hampel (1971) sense. However, any member of the σ_J family can be suitably trimmed to provide such robustness.

3. TWO SAMPLE TESTS

In normal samples the distribution of $\hat{\Delta}$ can be closely approximated by a χ, Ramasubban (1956), or by a Pearson type III curve, Barnett, Mullen, and Saw (1967). For other parent distributions the distribution of $\hat{\Delta}$ is unknown and not easily obtained. Nevertheless, the asymptotic variance of $\hat{\Delta}$ is stabilized within scale families by the log transformation, and Miller (1968) has shown how to use the jackknife, coupled with the log transformation, to provide approximate t statistics. We shall take this approach.

For a sample $X_1,\ldots,X_n$ and a scale statistic $\hat{\theta}$ we compute $\log\hat{\theta}$ and $\log\hat{\theta}_{-i}$, i=1,n, where $\hat{\theta}_{-i}$ is the statistic calculated with the ith observation missing. We form the pseudo-observations $\tilde{\theta}_i = n\log\hat{\theta} - (n-1)\log\hat{\theta}_{-i}$, the jackknife estimate of $\log\hat{\theta}$

$$\tilde{\theta}. = n\log\hat{\theta} - \frac{(n-1)}{n}\sum_{i=1}^{n}\log\hat{\theta}_{-i} ,$$

and the jackknife estimate of variance of $\tilde{\theta}.$

$$\hat{\sigma}(\tilde{\theta}.) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{\theta}_i - \tilde{\theta}.)^2 .$$

Then $(\tilde{\theta}. - \log\theta)/\{\hat{\sigma}(\tilde{\theta}.)\}^{\frac{1}{2}}$ is approximately t distributed with n-1 degrees of freedom. For two samples of size $n_1 = n_2 = n$ we compute $\tilde{\theta}_{.1}$, $\tilde{\theta}_{.2}$, $\hat{\sigma}(\tilde{\theta}_{.1})$, $\hat{\sigma}(\tilde{\theta}_{.2})$, and form our test statistic

$$T = \frac{(\tilde{\theta}_{.1} - \tilde{\theta}_{.2})}{\{\hat{\sigma}(\tilde{\theta}_{.1}) + \hat{\sigma}(\tilde{\theta}_{.2})\}^{\frac{1}{2}}}$$

which should be approximately t distributed with 2(n-1) degrees of freedom under the null hypothesis of equal scale.

Using the McGill "Super-Duper" random number generator, 1000 pairs of independent samples of size $n_1 = n_2 = 10$ and $n_1 = n_2 = 25$ were generated for selected families. The statistic T was calculated for the mean difference $\hat{\Delta}$, the usual sample variance s^2, the 1.5th power deviation $n^{-1}\Sigma_{i=1}^{n}|X_i - \bar{X}|^{1.5}$, and for the sample mean deviation $n^{-1}\Sigma_{i=1}^{n}|X_i - \bar{X}|$. Each was checked against the .05 and .01 percentage points of a t distribution, and the classical F test was also performed. Tables 4 and 5 list the empirical powers for ratios of variances $\sigma_1^2/\sigma_2^2 = 1,2,4,6,10$. The tables are intended to resemble Tables I and II of Miller (1968) to facilitate comparisons.

The F test shows its usual sensitivity to nonnormality and we will not mention it further. In Table 4 we observe that all tests are conservative for the uniform, with powers in the approximate order $s^2 > \hat{\Delta} >$ 1.5th deviation > mean deviation. For the normal s^2 dominates, but little is lost by using $\hat{\Delta}$ or the 1.5th deviation. As somewhat of a surprise s^2 continues to do well for the logistic with the 1.5th deviation slightly beating $\hat{\Delta}$, and all three dominating the mean deviation. For the Laplace s^2 tends to be too liberal, and a very slight ordering appears among the remaining three, 1.5th deviation > $\hat{\Delta}$ > mean deviation. Perhaps, if the median had been used rather than $\bar{X}$, the mean deviation would have performed better. For the exponential all tests are too liberal with s^2 the worst of the four; $\hat{\Delta}$ and the 1.5th deviation are indistinguishable

Table 4. Empirical powers for the two sample tests of scale; $n_1 = n_2 = 10$.

$\sigma_2^1/\sigma_2^2 =$	α = .05					α = .01				
	1	2	4	6	10	1	2	4	6	10
Uniform										
F	.016	.186	.748	.922	.991	.003	.037	.347	.686	.928
Sample variance	.023	.286	.783	.927	.985	.005	.080	.476	.717	.916
Sample 1.5th power	.021	.257	.726	.895	.977	.003	.060	.392	.659	.871
Sample mean deviation	.031	.225	.621	.809	.929	.004	.052	.320	.537	.774
Gini's mean difference	.021	.261	.749	.914	.979	.005	.066	.423	.686	.896
Normal										
F	.053	.265	.672	.844	.955	.011	.079	.379	.625	.854
Sample variance	.045	.194	.550	.734	.893	.008	.068	.232	.445	.676
Sample 1.5th power	.040	.185	.553	.738	.895	.006	.065	.223	.428	.680
Sample mean deviation	.044	.187	.500	.709	.863	.012	.063	.221	.387	.634
Gini's mean difference	.038	.183	.539	.741	.895	.006	.062	.224	.426	.667
Logistic										
F	.090	.325	.650	.798	.935	.028	.135	.434	.614	.808
Sample variance	.053	.217	.505	.664	.811	.009	.081	.246	.404	.593
Sample 1.5th power	.041	.203	.499	.660	.819	.009	.058	.231	.394	.587
Sample mean deviation	.036	.194	.470	.641	.797	.009	.046	.210	.363	.562
Gini's mean difference	.040	.204	.493	.655	.811	.007	.060	.224	.382	.576
Laplace										
F	.146	.359	.643	.780	.897	.064	.201	.441	.612	.783
Sample variance	.069	.201	.423	.549	.716	.025	.079	.204	.310	.460
Sample 1.5th power	.060	.188	.420	.567	.722	.018	.064	.190	.305	.468
Sample mean deviation	.060	.188	.411	.554	.720	.012	.054	.185	.294	.452
Gini's mean difference	.059	.182	.400	.554	.723	.018	.066	.179	.297	.448
Exponential										
F	.185	.361	.623	.734	.860	.093	.245	.427	.582	.744
Sample variance	.094	.219	.388	.496	.630	.030	.102	.213	.294	.420
Sample 1.5th power	.076	.208	.368	.482	.626	.021	.085	.193	.280	.395
Sample mean deviation	.069	.195	.354	.462	.597	.017	.065	.171	.267	.365
Gini's mean difference	.076	.196	.370	.489	.641	.019	.073	.188	.282	.399

Table 5. Empirical powers for the two sample tests of scale; $n_1 = n_2 = 25$.

	α = .05					α = .01				
$\sigma_2^1/\sigma_2^2 =$	1	2	4	6	10	1	2	4	6	10
					Normal					
F	.066	.542	.962	.995	1.000	.012	.273	.860	.981	.999
Sample variance	.056	.483	.938	.993	1.000	.013	.250	.804	.950	.997
Sample 1.5th power	.051	.479	.937	.994	1.000	.010	.234	.800	.955	.997
Sample mean deviation	.055	.450	.920	.987	.999	.011	.203	.751	.937	.995
Gini's mean difference	.052	.477	.939	.993	.999	.010	.230	.798	.954	.998
					Logistic					
F	.103	.493	.929	.990	1.000	.023	.276	.803	.959	.996
Sample variance	.053	.381	.843	.956	.989	.010	.172	.626	.843	.961
Sample 1.5th power	.045	.375	.874	.972	.998	.008	.165	.626	.881	.983
Sample mean deviation	.044	.375	.864	.971	.999	.006	.155	.609	.877	.985
Gini's mean difference	.041	.377	.873	.974	.998	.008	.164	.630	.880	.987
					Laplace					
F	.137	.515	.891	.971	.997	.062	.328	.786	.921	.989
Sample variance	.058	.312	.733	.881	.963	.011	.130	.465	.688	.877
Sample 1.5th power	.051	.311	.750	.903	.983	.006	.124	.474	.724	.906
Sample mean deviation	.048	.307	.756	.914	.985	.009	.115	.475	.728	.927
Gini's mean difference	.046	.301	.762	.910	.984	.005	.116	.476	.728	.915

but both dominate the mean deviation. The same pattern emerges in Table 5 for $n_1 = n_2 = 25$ except that s^2 is no longer a winner for the logistic, and $\hat{\Delta}$ and the 1.5th deviation are virtually indistinguishable for the logistic and Laplace as well as for the normal.

4. CONCLUDING REMARKS

Gini's mean difference has high asymptotic efficiency for scale estimation in symmetric distributions ranging from uniform to Laplace. Indications are that it will perform adequately in skewed situations as well. Two sample tests for scale can easily be based on log $\hat{\Delta}$ and tend to perform as well as those based on log 1.5th deviation.

REFERENCES

Barnett, F. C., Mullen, K. and Saw, J. G. (1967). Linear estimates of a population scale parameter. <u>Biometrika</u> <u>54</u>, 551-4.

Bickel, P. J. and Lehmann, E. L. (1976). Descriptive statistics for nonparametric models. III. Dispersion. <u>Ann</u>. <u>Statist</u>. <u>4</u>, 1139-58.

D'Agostino, R. B. (1970). Linear estimation of the normal distribution standard deviation. <u>Am</u>. <u>Statis</u>. <u>24</u>, 14-5.

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples, <u>Biometrika</u> <u>58</u>, 341-8.

David, H. A. (1968). Gini's mean difference rediscovered. <u>Biometrika</u> <u>55</u>, 573-5.

Downton, F. (1966). Linear estimates with polynomial coefficients. <u>Biometrika</u> <u>53</u>, 129-41.

Hampel, F. R. (1971). A general qualitative definition of robustness. <u>Ann</u>. <u>Math</u>. <u>Statis</u>. <u>42</u>, 1887-96

Johnson, N. L. and Kotz, S. (1970). <u>Continuous</u> <u>Univariate</u> <u>Distributions-1</u>. Boston: Houghton Mifflin.

Katti, S. K. (1960). Moments of the absolute difference and absolute deviation of discrete distributions. Ann. Math. Statis. 31, 78-85.

Miller, R. G. (1968). Jackknifing variances. Ann. Math. Statist. 39, 567-82.

Nair, U. S. (1936). The standard error of Gini's mean difference. Biometrika 28, 428-36.

Ramasuban, T. A. (1956). A χ-approximation to Gini's mean difference. J. Indian Soc. Agric. Statist. 8, 116-21.

Sarhan, A. E. (1954). Estimation of the mean and standard deviation by order statistics. Ann. Math. Statist. 25, 317-28.

Sarhan, A. E. (1955). Estimation of the mean and standard deviation by order statistics. II. Ann. Math. Statist. 26, 505-11.

von Andrae (1872). Ueber die Bestimmung des wahrscheinlichen Fehlers durch die gegeben Differenzen vom gleich genaven Beobachtungen einer Unbekannten. Aston. Nach. 79, 257-72.

Wainer, H. and Thissen, D. (1976). Three steps toward robust regression. Psychometrika 41, 9-34.