Psychol. Belg.
1983, XXIII-1, 73-83.

University of Ghent
Laboratory for Experimental, Differential and Developmental Psychology

SOME ROBUST STATISTICS FOR PSYCHOLOGISTS

ANDRÉ VANDIERENDONCK & GEERT DE SOETE*
**Aspirant N.F.W.O.*

The present paper discusses some robust statistics for psychologists. In a Monte Carlo study the efficiency of some robust alternatives to the mean and the standard deviation are studied in samples of varying sizes and varying degrees of deviation from normality. The median and the interquartile range appear to be insufficiently efficient under the normal model, while Gini's difference and a related estimator of the population mean seem to be not only robust under deviations from normality but also fairly efficient in normal samples. Furthermore, the usefulness of a specific robust statistic for estimating the population product-moment correlation is discussed. The article stresses the importance of robust statistics that are routinely applicable and quite efficient under conditions of normality. Some suggestions for statistical inferences based on these statistics are offered.

Statistical estimation and inference are based on the one hand on the observations and on the other hand on the assumed statistical model. If we want to estimate a population parameter or if we want to make inferences about population parameters, we need besides data some sort of statistical model. More than a century ago, Gauss (quoted in Huber, 1972), recognizing this fact, introduced the normal distribution as a statistical model because under this model the sample mean is an optimal estimator of the population mean. Since then the normal distribution and the associated method of least-squares have played a central role in classical statistics. Whenever the data *are* normally distributed, classical statistics provide the researcher with handy and efficient tools for statistical estimation and inference.

In practice, however, data are only *approximately* normally distributed. More specifically, data sets in the behavioral sciences often contain more outliers than one would expect on the basis of the normal distribution (cf. e.g., Wainer, 1976). This implies that the underlying distribution seems to have somewhat longer tails than the normal distribution. Are classical estimators such as the sample mean still efficient in such a situation? Unfortunately, it turns out that this is not the case. Therefore, two decades ago statisticians started

The authors are indebted to Prof. W. De Coster for providing the necessary computational facilities at the Laboratory for Experimental, Differential and Developmental Psychology of the University of Ghent.

developing methods for estimation and inference that are *relatively insensitive to small deviations from the assumed statistical model.* These methods are referred to as *robust statistics* (for a survey, see Huber, 1972, 1981).

In this paper we present statistics for estimating the population mean, the population standard deviation, and the population product-moment correlation, which are not only robust in the sense discussed above, but which are also 1. in principle routinely applicable and 2. sufficiently efficient when the data *are* normally distributed.

ROBUST ESTIMATION OF LOCATION AND SCALE

Wainer and Thissen (1976) studied a robust estimator of scale, s*, which is related to Gini's difference statistic (Gini, 1912). Let $\underline{x}$ be a vector of n ordered observations, such that

$$x_1 \leqslant x_2 \leqslant \ldots \leqslant x_i \leqslant \ldots \leqslant x_n.$$

The i'th gap is defined as

$$g_i = x_{i+1} - x_i, \; i = 1, \ldots, n-1. \tag{1}$$

Generally, the data tend to cluster around a central value or typical location, and hence, the gaps tend to be larger the closer i is to 1 or $n-1$. In order to obtain a robust estimator, these extreme values should be weighted less than the central ones. Therefore, the gaps g_i are combined with a set of weights w_i :

$$w_i = i(n-i), \; i = 1, \ldots, n-1, \tag{2}$$

which are symmetric and approximately normally distributed.

The estimator of scale, s*, is then expressed as

$$s^* = \frac{\sqrt{\pi}}{n(n-1)} \sum_{i=1}^{n-1} w_i g_i, \tag{3}$$

where $\sqrt{\pi}/n(n-1)$ is introduced to make s* an unbiased estimator of σ, the population standard deviation, when the data are sampled from a normal distribution. I.e., $\mathcal{E}(s^*) = \sigma$ under the normal model.

With some algebraic manipulation it is easy to show that

$$s^* = \frac{\sqrt{\pi}}{n(n-1)} \sum_{i=1}^{n} (2i - n - 1)\, x_i, \tag{4}$$

a form which is computationally more efficient.

The estimator s* has some attractive properties :

$$s^*(\underline{x} + a) = s^*(\underline{x}), \tag{5}$$

and

$$s^*(a\underline{x}) = a[s^*(\underline{x})]. \tag{6}$$

Equation (5) states that s^*, is invariant under translations of $\underline{x}$, while equation (6) means that multiplying all data values by the same constant, has a similar effect on s^*. The latter feature implies that the observations can be standardized to yield $s^* = 1$:

$$s^*[\underline{x}/s^*(\underline{x})] = s^*(\underline{x})/s^*(\underline{x}) = 1.$$

Downton (1966) has shown that s^* has an asymptotic efficiency of 98% relative to the usual estimator of σ, s :

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - m)^2}{n - 1}}$$

where

$$m = \sum_{i=1}^{n} x_i/n.$$

Wainer and Thissen (1976) demonstrated in a small-scale Monte Carlo study that s^* has the same properties for finite samples. They showed not only that s^* is almost as efficient as s when the data are normally distributed, but also that s^* is much more efficient than s when the underlying distribution has longer tails than the normal distribution.

Tukey (1977) proposed the use of the interquartile range (IQR) as a rough-and-ready robust alternative to s. Therefore, we decided to compare the bias and efficiency of s, s^* and IQR in a more extensive Monte Carlo study, using not only more replications but also a larger design than Wainer and Thissen (1976).

In addition to estimating the variability, investigators are also (or even more) interested in estimating the central value of a set of observations. Since the classical estimator of location, m, is rather sensitive to outliers, a more robust estimator of the population mean, μ, is called for.

An estimator of location which is naturally related to IQR is the median, md. However, it is known that the median is not very efficient compared to the sample mean, when the data are normal. Asymptotically, its relative efficiency is about $2/\pi$ (Mosteller and Tukey, 1968). The loss of efficiency in case of normality is too large to be acceptable. On the other hand, it is instructive to investigate the behavior of md when the assumption of normality is not completely satisfied.

Well-known procedures for arriving at robust estimators of location are trimming and winsorizing (cf. e.g., Wainer, 1976). A trimmed mean is defined as

$$m_t = \sum_{i=1}^{n} v_i x_i, \quad i = 1, \ldots, n, \tag{7}$$

with

$$v_i = 0, \text{ if } i < k \text{ or } i > n - k + 1$$

$$v_i = 1, \text{ otherwise}$$

where k is an appropriately chosen value between 0 and n/2. Equation (7) implies that the (k – 1) smallest and the (k – 1) largest observations are completely ignored in estimating μ. This is a severe treatment of extreme observations. Moreover, the choice of k is rather arbitrary. Therefore, trimmed means cannot be considered to be routinely applicable.

The technique of winsorizing is related to trimming. Extreme values are not completely discarded from the sample, but the g most extreme values on both sides are replaced by the (g+ 1)'th and the (n – g)'th value respectively. Although this treatment of extreme values is less drastic than in trimming, its application requires a decision about the size of g.

We rather prefer an estimator which is a weighted average of the data, with weights which have an approximately Gaussian distribution :

$$m^* = \sum_{i=1}^{n} w_i x_i \Big/ \sum_{i=1}^{n} w_i \qquad (8)$$

where $w_i = i(n+1-i)$. In fact the same weights are utilized as those applied to the gaps, g_i, in the definition of s* (equations 1-3)[1].

MONTE CARLO STUDY

In an extensive Monte Carlo simulation we compared the estimators of location m, m* and md, and the estimators of scale s, s* and IQR. The purpose of this study was to obtain relevant information about the behavior of these statistics in samples varying in size and underlying distribution. As already mentioned, data usually deviate from normality in having fatter tails than a normal distribution. This situation can be simulated by sampling from a contaminated normal distribution. Such samples are constructed by drawing $100(1-p)\%$ of the data from a standard normal distribution, N(0,1), and the remaining part ($100p\%$) from a normal population with the same mean but with a larger variance, usually N(0,9).

The design of the study consisted of a factorial combination of sample size (n = 10, 20, 50 or 100) and degree of contamination (0, 5, 10, 20 or 50 percent drawn from N(0,9)). There were 1000 replications per cell. Normally distributed random numbers were obtained using Brent's (1974) algorithm which was fed with uniformly distributed

---

[1] Remark that the weights applied in m* and s* are statistically independent. I.e.,

$$\sum_{i=1}^{n} i(n+1-i)(2i-n-1) = 0.$$

To prove this equality, define $w_i = i(n+1-i)$ and $v_i = (2i-n-1)$. Observe that for any j, $1 \leq j \leq n/2$, $w_j = w_{n+1-j}$, and $v_j = -V_{n+1-j}$, so that $w_j v_j + W_{n+1-j} v_{n+1-j} = 0$.

numbers generated by Schrage's (1979) portable random number generator.

For each sample, the six statistics discussed above were computed. The mean and the variance of each statistic over the 1000 replications were used to estimate the mean and the variance of the sampling distribution of the statistic.

Finally, in each condition, the efficiency of m* and md relative to m and s* and IQR relative to s were estimated. The estimated relative efficiency of a statistic T_1 in comparison to another statistic T_2 is defined as Var (T_2)/Var(T_1).

Table 1 summarizes the results. Inspection of the table shows that none of the estimators of μ is biased. When contamination is present, both m* and md are more efficient than m. However, when there is no or very little contamination, md loses – as we expected – much of its efficiency.

The results on the three estimators of σ show that when the degree of contamination is moderate to high, both s* and IQR are more efficient than s. On the contrary, when the contamination is small or nonexistent, IQR is obviously inefficient.

Taken together, these findings confirm and extend those of Wainer and Thissen (1976) concerning the robustness of s*. In addition the present study shows that m*, a measure of location related to s*, is also very robust, and is efficient when the population is normal: 97 to 99%. On the other hand, md and IQR, although very efficient when the degree of contamination is high, appear to be less suitable alternatives to m and s because of their inefficiency in normal samples.

The conclusions to be drawn from this study are obvious: *Given that many real data sets deviate substantially from normality in having longer tails, m* and s* are more useful statistics than m and s, respectively, the increased computational cost notwithstanding.* Whereas for small and moderately large samples the cost of computing m and s is negligible, this is not true for m* and s*. This is mainly due to the fact that m* and s* require the data to be sorted. However, as CPU time has become cheaper, this should not be a real objection anymore.

ROBUST CORRELATION

Another statistic which is frequently used by psychologists besides m and s, is the sample product-moment correlation coefficient r. Unfortunately, r is greatly influenced by bivariate outliers, just as m and s are very sensitive to univariate outliers. Bivariate outliers can be detected by means of sample influence functions and influence-enhanced scatterplots. A sample influence function is a function which describes the impact of each data point on the value of an estimator. Devlin, Gnanadesikan and Kettenring (1975) suggested to define the influence of the i'th pair of observations (x_i, y_i) on r as follows

$$I\_(x_i, y_i; r) = (n-1)(r - r_{(-i)})$$

TAB. 1. BEHAVIOR OF m, m*, md, s, s*, AND IQR UNDER VARYING DEGREES OF CONTAMINATION AND SAMPLE SIZES.

n	%	measures of location				
		m	m*	md	E(m*)	E(md)
10	0	.004	.004	.011	0.96	0.69
	5	−.002	−.002	−.004	1.14	0.96
	10	−.013	−.012	−.009	1.27	1.13
	20	−.004	−.001	−.003	1.33	1.29
	50	−.010	−.014	−.011	1.25	1.52
20	0	.002	.002	.005	0.96	0.68
	5	.004	.003	.003	1.17	0.88
	10	.004	.005	.004	1.32	1.07
	20	.003	.003	−.005	1.43	1.27
	50	−.025	−.022	−.022	1.34	1.53
50	0	.000	.000	−.003	0.98	0.68
	5	.001	−.003	−.006	1.19	0.89
	10	−.007	−.006	−.003	1.34	1.00
	20	−.008	−.007	−.007	1.54	1.35
	50	.003	.000	−.002	1.37	1.39
100	0	−.002	−.001	−.002	0.96	0.65
	5	.001	.002	.003	1.23	0.87
	10	−.003	−.003	.001	1.33	0.94
	20	−.009	−.008	−.009	1.56	1.29
	50	.011	.007	.000	1.35	1.49

n	%	measures of scale				
		s	s*	IQR	E(s*)	E(IQR)
10	0	0.97	1.00	1.22	0.93	0.40
	5	1.12	1.12	1.29	1.30	1.03
	10	1.27	1.25	1.37	1.40	1.12
	20	1.54	1.49	1.52	1.33	1.22
	50	2.14	2.13	2.30	1.10	0.74
20	0	0.99	1.00	0.99	0.96	0.44
	5	1.14	1.12	1.03	1.56	1.24
	10	1.30	1.25	1.09	1.65	1.64
	20	1.57	1.48	1.17	1.49	1.94
	50	2.21	2.14	1.67	1.14	1.14
50	0	1.00	1.00	1.04	0.98	0.40
	5	1.17	1.12	1.08	1.75	0.99
	10	1.34	1.25	1.14	1.87	1.63
	20	1.59	1.47	1.24	1.74	1.79
	50	2.21	2.11	1.70	1.17	1.07
100	0	1.00	1.00	1.00	0.97	0.35
	5	1.18	1.13	1.04	1.87	1.19
	10	1.33	1.24	1.08	1.79	1.51
	20	1.61	1.48	1.18	1.64	1.57
	50	2.21	2.10	1.58	1.18	1.07

Note. The symbol n indicates the sample size, % refers to the percentage of each sample drawn from a contaminating population, and E is the relative efficiency of the statistic.

where $r_{(-i)}$ is the sample product-moment correlation obtained after deleting (x_i, y_i). Recently, Thissen, Baker and Wainer (1981) discussed a variety of interesting ways for plotting the values of this influence function. Such plots, referred to as influence-enhanced scatterplots, are scatterplots in which each data point is supplemented with a graphical indication of its influence, for instance by means of a directed line segment whose length is proportional to the magnitude of $I_{-}(x_i, y_i; r)$.

In order to illustrate the non-robustness of r, we calculated r for a data set containing an undeniable bivariate outlier. The data, borrowed from Efron (1979), consist of the average LSAT (Law School Admission Test) score and the average GPA (undergraduate Grade Point Average) of entering students in 15 American law schools. The value of r amounts to 0.776 and an influence-enhanced scatterplot is presented in Figure 1. In this figure, an upward line indicates a positive influence

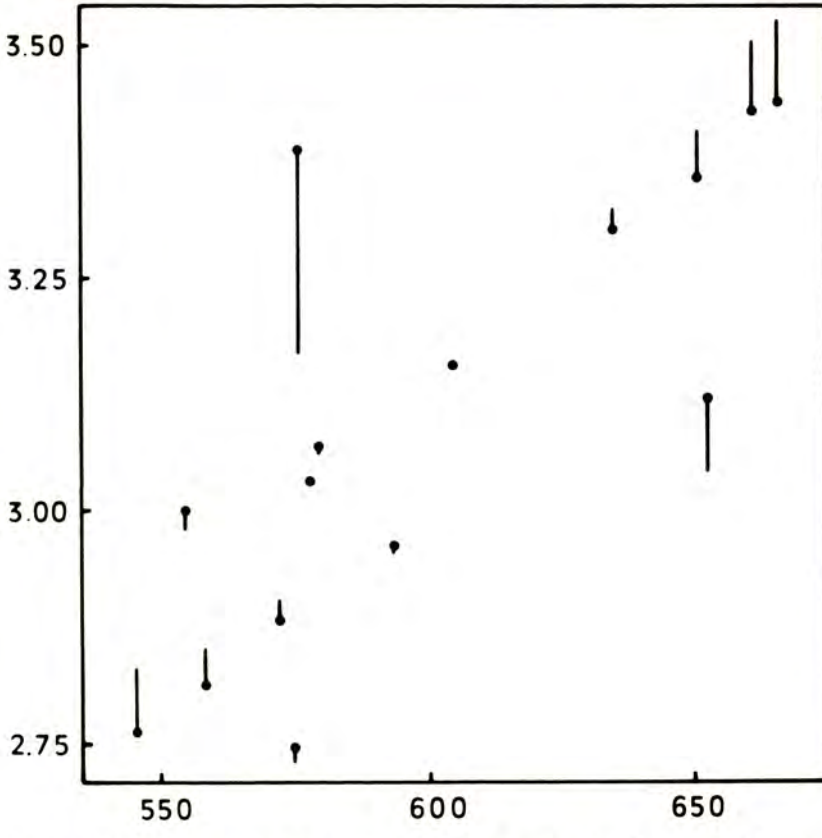

FIG. 1. INFLUENCE-ENHANCED SCATTERPLOT OF r FOR THE EFRON (1979) DATA. The LSAT score is plotted on the abscissa and the GPA on the ordinate.

while a negative influence is indicated by a downward line segment. The length of the line segments is proportional to the magnitude of the influence. It is apparent from Figure 1 that some points influence r more than others. This is typical for a non-robust estimator.

Because of the non-robustness of r, we should look for another estimator of the population correlation coefficient, which is less sensitive to bivariate outliers. An interesting alternative to r, denoted r*, has been discussed by Wainer and Thissen (1976). The robust estimator r*, attributed by Wainer and Thissen (1976) to John W. Tukey, is defined by

$$r^* = (1/4)\{[s^*(\tilde{\underline{x}} + \tilde{\underline{y}})]^2 - [s^*(\tilde{\underline{x}} - \tilde{\underline{y}})]^2\}, \tag{10}$$

where

$$\tilde{\underline{x}} = \underline{x}/s^*(\underline{x})$$

$$\tilde{\underline{y}} = \underline{y}/s^*(\underline{y}).$$

Equation (10) is based on the fact that

$$\begin{aligned} &\mathrm{Var}(\underline{X} + \underline{Y}) - \mathrm{Var}(\underline{X} - \underline{Y}) = \\ &\mathrm{Var}(\underline{X}) + \mathrm{Var}(\underline{Y}) + 2\rho\sqrt{\mathrm{Var}(\underline{X})\mathrm{Var}(\underline{Y})} \\ &- \mathrm{Var}(\underline{X}) - \mathrm{Var}(\underline{Y}) + 2\rho\sqrt{\mathrm{Var}(\underline{X})\mathrm{Var}(\underline{Y})} = \\ &4\rho\sqrt{\mathrm{Var}(\underline{X})\mathrm{Var}(\underline{Y})} \end{aligned}$$

where $\underline{X}$ and $\underline{Y}$ are two random variables such that

$$\mathrm{Covar}(\underline{X}, \underline{Y}) = \rho\sqrt{\mathrm{Var}(\underline{X})\mathrm{Var}(\underline{Y})}.$$

When compared to r, this estimator is less influenced by outliers along the x- and y- axes, but also by outliers along the 45° and −45° lines. Wainer and Thissen (1976) further studied the behavior of r*, relative to r and the Spearman rank correlation r_s in a Monte Carlo experiment. They found that both r* and r_s are robust estimators of the population correlation coefficient *and* that both are quite efficient even when the data are sampled from an uncontaminated normal distribution. Especially when regression weights are needed, r* should be preferred to r_s, because of the existence of a related robust estimator of scale (viz., s*).

When applied to the data in Figure 1, r* amounts to 0.821. An influence-enhanced scatterplot of $I\_(x_i, y_i; r^*)$ is presented in Figure 2. Comparing this figure with Figure 1 reveals that although the average magnitude of $I\_(x_i, y_i; r)$ is approximately equal to the average magnitude of $I\_(x_i, y_i; r^*)$, the variances of $I\_(x_i, y_i; r)$ and $I\text{–}(x_i, y_i; r^*)$ (0.175 and 0.090 respectively) are quite different, suggesting that r* is more evenly influenced by all observations than r.

When more than two variables are involved, statistical analysis is often based on an estimate of the population correlation matrix. A

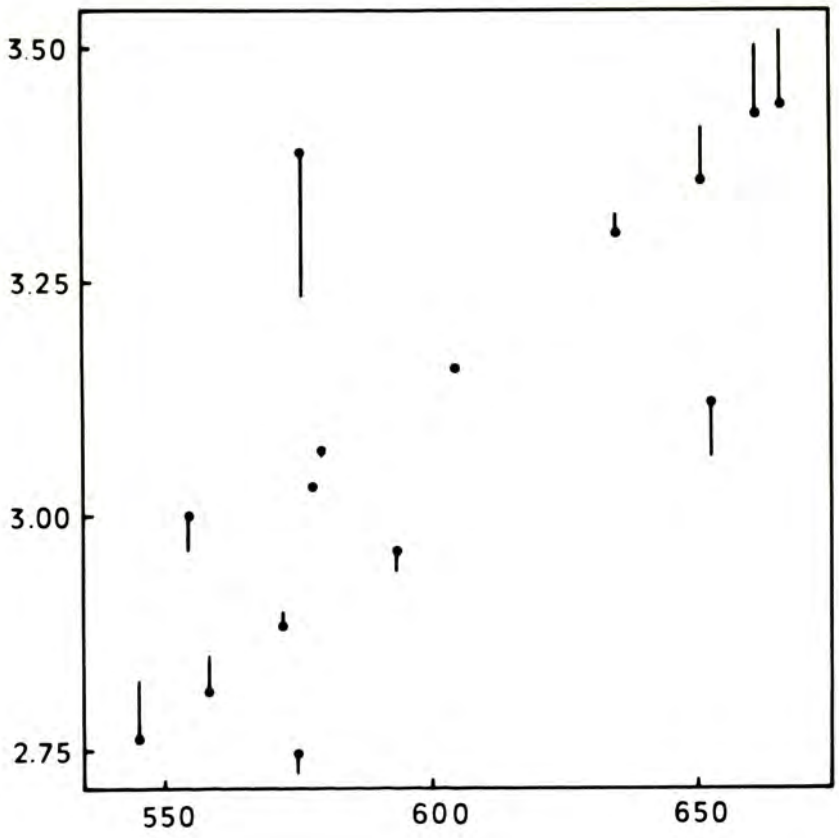

FIG. 2. INFLUENCE-ENHANCED SCATTERPLOT OF r* FOR THE EFRON (1979) DATA. The LSAT score is plotted on the abscissa and the GPA on the ordinate.

robust estimate can be obtained by using r* instead of r. There is however one problem. A correlation matrix is required to be non-negative definite. The non-negative definiteness of a correlation matrix can be thought of as a high-dimensional analogue of the property that a single correlation coefficient lies between -1 and $+1$. When r* is used to compute a robust estimate R*, the estimate is not necessarily non-negative definite. One way to solve the problem is to shrink R* nonlinearly towards the identity matrix when R* is not non-negative definite (cf. Devlin et al., 1975). I.e., each element r_{ij}* of R* is replaced by a nonlinear transformation of it, $g(r_{ij}$*). One particular kind of nonlinear transformation has been proposed by Devlin et al. (1975):

$$g(r_{ij}^*) = \begin{cases} \tanh[\tanh^{-1}(r_{ij}^* + \varepsilon)] \text{ if } r_{ij}^* < -\tanh(\varepsilon) \\ 0 \text{ if } |r_{ij}^*| \leqslant \tanh(\varepsilon) \\ \tanh[\tanh^{-1}(r_{ij}^* - \varepsilon)] \text{ if } r_{ij}^* > \tanh(\varepsilon) \end{cases} \tag{11}$$

where ε is some small positive constant (e.g., 0.05). This transformation can be repeatedly applied on the adjusted correlation matrix until the latter has no negative eigenvalues anymore. The final adjusted R^* can then be used to perform a variety of robust statistical analyses, such as robust multiple regression, robust principal components analysis, and robust factor analysis.

CONCLUSION

In this paper we presented robust alternatives to the usual least-squares estimators of the population mean, the population standard deviation, and the population product-moment correlation. We deliberately chose estimators which are 1. in principle routinely applicable and 2. sufficiently efficient when the data are normally distributed. Although we primarily focused on robust estimation, it should be clear that these robust estimators can be used to arrive at robust statistical inference, for instance by means of jackknifing and bootstrapping (cf. e.g., De Soete and Vandierendonck, 1982) or (approximate) randomization tests (cf. e.g., De Soete, 1982).

REFERENCES

BRENT, R.P. Algorithm 488. A Gaussian pseudo-random number generator. *Communications of the ACM*, 1974, *17*, 704-706.

DE SOETE, G. A note on the use of parametric versus nonparametric tests for comparing means. *Tijdschrift voor Onderwijsresearch*, 1982, *7*, 182-184.

DE SOETE, G., & VANDIERENDONCK, A. On the use of the jackknife and the bootstrap for estimating a confidence interval for the product-moment correlation coefficient. *Psychologica Belgica*, 1982, *22*, 87-97.

DEVLIN, S.J., GNANADESIKAN, R., & KETTENRING, J.R. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 1975, *62*, 531-545.

DOWNTON, F. Linear estimates with polynomial coefficients. *Biometrika*, 1966, *53*, 129-141.

EFRON, B. Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, 1979, *21*, 460-480.

GINI, C. Variabilità e mutabilità, contributo allo studio delle distribuzioni e relazioni statistiche. *Studi-Economico-Giuridici della R. Università di Cagliari*, 1912.

HUBER, P.J. Robust statistics: A review. *Annals of Mathematical Statistics*, 1972, *43*, 1041-1067.

HUBER, P.J. *Robust Statistics*. New York: Wiley, 1981.

MOSTELLER, F., & TUKEY, J.W. Data analysis, including statistics. In G. LINDZEY & E. ARONSON (Eds.), *Handbook of social psychology* (2nd edition). Reading, Massachusetts: Addison-Wesley, 1968.

SCHRAGE, L. A more portable Fortran random number generator. *ACM Transactions on Mathematical Software*, 1979, *5*, 132-138.

THISSEN, D., BAKER, L., & WAINER, H. Influence-enhanced scatterplots. *Psychological Bulletin*, 1981, *90*, 179-184.

TUKEY, J.W. *Exploratory data analysis*. Reading, Massachusetts: Addison-Wesley, 1977.

WAINER, H. Robust statistics: A survey and some prescriptions. *Journal of Educational Statistics*, 1976, *1*, 285-312.

WAINER, H., & THISSEN, D. Three steps towards robust regression. *Psychometrika*, 1976, *41*, 9-34.

University of Ghent
H. Dunantlaan 2
9000 Ghent

Received May 1982