DISAGREEMENT: ESTIMATION OF RELATIVE BIAS OR DISCREPANCY RATE

by

PING HANG MA

B.Sc., The University of British Columbia, 1984

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(The Department of Statistics)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 1987

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date Sept. 28, 1987.

ABSTRACT

Not only basic research in sciences, but also medicine, law, and manufacturing need statistical techniques, including graphics, to assess disagreement. For some items or individuals $i = 1, 2, \cdots, n$ suppose that pairs (X_i, Y_i) denote each item's measurements by two distinct methods or by two observers, or X_i and Y_i may be initial and repeat measurement scores, with discrepancy $D_i = X_i - Y_i$. Disagreement may be characterized by location and scale parameters of discrepancy distributions.

The present work primarily addresses estimation of central tendency — relative bias or median discrepancy (or discrepancy *rate* in some instances). Most previous literature on "agreement" or "reliability" instead concerns X, Y correlation, which can be regarded as the complement of discrepancy variance. (There is ambiguity or confusion about concepts of "reliability" in the literature of various applications.)

Discrepancies $D_1, D_2, \cdots, D_n$ in practice often violate assumptions of standard statistical models and methods that have been commonly applied in studies of agreement. In particular, both X_i and Y_i generally incorporate measurement errors. Further, these two measurement error distributions for the i^{th} item need not be the same; and both distributions could depend on the magnitude μ_i of the item being measured. Hence, for example, discrepancy D_i could have variance proportional to the size of the item;

and in general $D_1, D_2, \cdots, D_n$ are *not* identically distributed. Finally, the selection of items $i = 1, 2, \cdots, n$ often is not random.

To estimate median discrepancy, we consider nonparametric confidence intervals corresponding to Student t test, sign test, Wilcoxon signed rank test, or other permutation tests. Several criteria are developed to compare the performance of one procedure relative to another, including expected ratio of confidence interval lengths (related to Pitman asymptotic relative efficiency of tests) and relative variability of interval lengths. Theoretical calculations and Monte Carlo simulation results suggest different procedural preferences for random sampling from different distributions.

For discrepancies distributed non-identically, but symmetrically about a common median value, mixture sampling is used as an approximate model. This approach is related to a "random walk" (rather than random sample) model of $D_1, D_2, \cdots, D_n$ proposed particularly for discrepancies between counting processes.

We also emphasize graphic methods, especially plots of difference of $Y - X$ versus average $(X + Y)/2$, for exploratory analysis of discrepancy data and to choose appropriate statistical models and numerical methods.

Various data sets are analyzed as examples of the methodology.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF DATA SETS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank Dr. Ned Glick for his guidance, assistance and encouragement in producing this thesis, as well as suggesting the topic. I am indebted to Dr. Jonathan Berkowitz for his useful comments and careful reading of this work, in addition to his concern throughout the years. Encouragement and support from other members of the Department of Statistics also are gratefully appreciated.

I thank attorneys James W. Peters, Brian J. Wallace, and E. J. Gouge for discussions of the logging data in Example 7.1, originally analyzed by Dr. Glick.

I owe my wife, May-Moon, and my son, Lok-Chun, for their patience and support, while I spent days away from home.

This work received financial support in part from Dr. Glick's forest industry consulting, as well as my teaching and research assistanceships in the Department of Statistics, University of British Columbia.

1. INTRODUCTION

Medicine, manufacturing, and research in sciences all require counting of items or measuring amounts of substances being studied. Therefore, it is not surprising that great effort and time are devoted to evaluating measurement methodologies, from intra-observer and inter-observer perspectives, and comparing distinct measurement methods. For instance, Beeler (1986) found that as many as one-third of all papers published in *American Journal of Clinical Pathology* are method comparison studies. Developments of new or "improved" methods may offer operational advantages, such as in speed, cost, convenience, etc. Consequently, there are many contexts in which the statistician requires techniques for comparing repeated or paired measurements or comparing one measurement process to another. This thesis is particularly concerned with measurement scales that are continuous or that permit integer values over a large range, although there is some consideration of ordinal categoric ratings.

Suppose $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ are the pairwise measurements made by two measuring methods or by two observers or at two distinct occasions on n items. For instance (X_i, Y_i) may be the counts of red blood cells in the i^{th} blood specimen by two cell counting devices or the finishing

times of the i^{th} racer recorded by two observers. (Other examples are given in Chapter 7.) Then define their *discrepancies* $D_1, D_2, \cdots, D_n$ by

$$D_i = X_i - Y_i, \qquad i = 1, 2, \cdots, n.$$

The term "discrepancy" or "disagreement" reflects the possibility of error in both X_i and Y_i observations: if μ_i denotes the "true" value associated with the i^{th} item, then represent the two measurements, respectively, as $X_i = \mu_i + \varepsilon_i$ and $Y_i = \mu_i + \delta_i$, where ε_i and δ_i are random errors (with respective biases $E\varepsilon_i$ and $E\delta_i$). Thus, no available measurement process is assumed to be absolutely accurate.

Agreement or disagreement between paired measurements can be characterized by at least two aspects — relative bias and reliability, or precision. "Relative bias" refers to the mean of the probability distribution of discrepancy D_i:

$$\begin{aligned} E(D_i) &= E(X_i - Y_i) \\ &= E(X_i - \mu_i) - E(Y_i - \mu_i) \\ &= E(\varepsilon_i) - E(\delta_i) \\ &= \text{bias of } X_i - \text{ bias of } Y_i \\ &= \text{relative bias of } X_i \text{ and } Y_i. \end{aligned}$$

Median or other central location parameter, rather than expectation of D_i, also may be of interest.

"Reliability" or "precision" or "reproducibility" usually refers to the predictability of one measurement, given the other. This issue essentially relates to the variance of discrepancy distribution — or, equivalently, to the correlation coefficient, $\rho(X_i, Y_i)$, of the joint distribution the X_i and Y_i measurements:

$$\begin{aligned} Var(D_i) &= Var(X_i - Y_i) \\ &= Var(X_i) + Var(Y_i) - 2Cov(X_i, Y_i) \\ &= Var(X_i) + Var(Y_i) - 2[\rho(X_i, Y_i)]\sqrt{Var(X_i)}\sqrt{Var(Y_i)}. \end{aligned}$$

Note that if $Var(X_i) = Var(Y_i) = \sigma_i^2$, as in several models considered below, then $Var(D_i) = 2\sigma_i^2[1 - \rho(X_i, Y_i)]$. But $Var(X_i)$ and $Var(Y_i)$, and hence variance of the discrepancy distribution, may be in any functional relationship with μ_i; the variances may be, for instance, proportional to μ_i or to μ_i^2, etc.

This thesis is mostly concerned with relative bias — assessment of central tendency for discrepancy distributions. Emphasis is on graphical methods and estimation procedures, including confidence intervals, for the mean or median of discrepancy distribution. Motivation comes partly from

data analysis (by Professor Ned Glick) in litigation where substantial dollar costs were claimed in proportion to an alleged relative bias or discrepancy rate in measuring quantities of wood; see Chapter 7.

There appears to be little statistical literature directly related to the relative bias aspect of agreement or disagreement (although there is much literature on "reliability" as noted in Chapter 2, and much general theory related to location parameter estimation and confidence interval issues, as discussed in later chapters). The most directly relevant material seems to be Altman and Bland (1983) and Bland and Altman (1986), and some responses to their works — partly published while this thesis was in progress.

2. SURVEY OF "AGREEMENT" IN RESEARCH LITERATURE

As discussed in the Introduction, there are at least two aspects of agreement or disagreement between measurements — namely, relative bias and reliability; or equivalently, location and scale parameters of the discrepancy distribution. There seems to be little explicit attention to relative bias in research literature; a notable exception is Bland and Altman (1986) using the graphical approach proposed by Altman and Bland (1983). There is much literature on reliability using techniques such as correlation, intraclass correlation coefficient, Cohen's kappa, linear regression, general linear model, etc. For review of statistical methods in reliability studies, see Landis and Koch (1975, Parts I and II). But many applications in reliability literature, discussed in this chapter, either misinterpret these techniques or are based on assumptions that may be invalid, for example, assuming that measurement errors have the same variance for different observers and for all items measured. Also the two aspects of agreement often are confused, for instance, in research literature of medicine or behavioural sciences. In particular, the term "agreement" often is used as a synonym for reliability, neglecting relative bias.

Altman and Bland (1983) suggested that such confusion may arise

"because virtually all introductory courses and textbooks in statistics are method-based rather than problem-based" — that is, correlation is a prominently used elementary method, but the problem(s) of assessing agreement may be nowhere mentioned. "A further reason for poor methodology is", according to Altman and Bland (1983), "the tendency for researchers to imitate what they see in other published papers". A related issue is that many discussions of "agreement" (or of "reliability") calculate some quantitative statistics(s) without clearly indicating why one should be interested, nor what may be the practical implications or applications.

Altman and Bland (1983) noted that relative bias is a more important issue than reliability in a published comparison of two methods for measuring systolic blood pressure. In another example, mentioned in the last chapter, substantial dollar costs were claimed in proportion to an alleged relative bias or discrepancy rate in counting and measuring volumes of wood.

2.1. CORRELATION

Some research workers tacitly — and incorrectly — assume that repeatability, usually evaluated in terms of high correlation or "significant" linear regression, implies low relative bias. Such fallacy is common in elementary

statistics and has been discussed, for instance, by Freedman, Pisani and Purves (1978). Considering the Skeels-Skodak study for intelligence scores of adopted children, their adoptive mothers, and their biologic mothers, Freedman, Pisani and Purves (1978, pp.139–141) noted that correlation may be stronger between children and their biologic mothers than between children and their adoptive mothers, but that the average score for these children could be much closer to the average of adoptive mothers. In terms of linear regression to predict children's scores from their biologic mothers' scores, the intercept may be large, although the slope is close to one and is "highly significant". The biologic mothers may "predict" their children's scores in the sense of explaining a large fraction of variation in the dependent variable; but the distribution of scores for the children could be systematically shifted with respect to the distribution of their biologic mothers' scores.

Correlation coefficient also has been incorrectly interpreted as a percentage of agreement. Cassidy, Triplett and LaDuca (1985) studied the Factor VIII inhibitors in blood, evaluating agreement between two measuring methods and between two laboratories. Because all their pairwise correlation coefficients are roughly equal to 0.9, the authors concluded that "these values indicate approximately 90% agreement for each comparison".

Other researchers seem to interpret *squared* correlation as an agreement scale. Because correlation coefficient usually is close to one in reliability studies, Rawles (1986) suggested squaring to "spread" out the "cramped" "meaningful range".

In the present context, both X and Y measurements are subject to (non-degenerate) errors; this implies that the expectation of the sample correlation coefficient always is less than one. This phenomenon sometimes is called "attenuated correlation". See Altman and Bland (1983) or Fleiss (1986, pp.3–4). Such components-of-variance perspective also shows that correlation depends on the mechanism by which objects or "items" are selected, and is not an intrinsic property of the measurement procedures. Indeed, in many cases, items to compare measurement methods or to assess agreement between ratings are not drawn by any random procedure, but arbitrarily or deliberately; and a great range of scores usually would lead to a high correlation, regardless of relative bias between measurements [Altman and Bland (1983)].

2.2. INTRACLASS CORRELATION COEFFICIENT

In literature of behavioural sciences and elsewhere, intraclass correlation

coefficient (ICC) has been outstandingly used to measure reliability or repeatability of measurement procedures, with two or more observations per item; see Gulliksen (1950), Ebel (1951), Guilford (1954), Haggard (1958), Hoyt and Krishnaiah (1960), Winer (1962), Hoffman (1963), and others.

Bartko (1966) showed that, for pairs, ICC and the usual Pearson correlation coefficient estimate the same parameter. He also showed that the ICC applies in a linear model in which "item" is a random effect; but using fixed effect data — items that are arbitrarily chosen — ICC does not resolve the problem of correlation depending on the item selection mechanism, as discussed in last section.

2.3. KAPPA

Cohen (1960) introduced a kappa statistic as a measure of inter-rater agreement for categoric data. Later Cohen (1968) generalized to a weighted kappa, which allows the relative seriousness of each disagreement to be quantified. For full discussion of kappa, see Chapter 13 of Fleiss (1981).

Suppose two raters (or measuring methods) separately classified n items on an L-point scale; the resulting data can be summarized in an $L \times L$ contingency table, or, equivalently, an array of observed proportions, such

that p_{ij} denotes the proportion of subjects classified into i^{th} category by the 1^{st} rater and into j^{th} category by the 2^{nd} rater. Since agreement requires raters to classify a given subject identically into the same category, one simple index of agreement is estimated by

$$p_0(\omega) = \sum_{i=1}^{L} \sum_{j=1}^{L} \omega_{ij} p_{ij},$$

where $\{\omega_{ij}, \quad i,j = 1,2,\cdots,L\}$ are a set of non-negative weights, assigned according to the seriousness of disagreement (and independently of the data actually collected).

Originally Cohen (1960) took $\omega_{ii} = 1$ (corresponding to agreement) and $\omega_{ij} = 0$ for $i \neq j$ (any disagreement), so that $p_0 = \sum_{i=1}^{L} p_{ii}$. In general, we require that

$$\omega_{ii} = 1,$$

$$0 \leq \omega_{ij} < 1 \text{ for } i \neq j, \text{ and}$$

$$\omega_{ij} = \omega_{ji}.$$

See Feldman, Klein, and Honingfeld (1972), and Cicchetti (1976) for different choices of ω_{ij}.

Cohen takes account of chance-expected agreement. If we assume independence between ratings by the two raters, the expected agreement

proportion is estimated by

$$p_e(\omega) = \sum_{i=1}^{L} \sum_{j=1}^{L} \omega_{ij} p_{i.} p_{.j},$$

where $p_{i.} = \sum_{k=1}^{L} p_{ik}$, and $p_{.j} = \sum_{k=1}^{L} p_{kj}$. Then $p_0(\omega) - p_e(\omega)$ represents the observed excess agreement beyond chance, and $1 - p_e(\omega)$ indicates the maximum possible excess agreement beyond chance. Cohen proposed a measure of agreement, adjusting for the agreement due to chance: the weighted kappa statistic is

$$\hat{\kappa}(\omega) = \frac{p_0(\omega) - p_e(\omega)}{1 - p_e(\omega)},$$

which ranges from $\frac{-p_e(\omega)}{1 - p_e(\omega)}$ to 1, with the lower value depending on the marginal distributions. Note that only for the special case where $p_e(\omega) = \frac{1}{2}$ does $\hat{\kappa}(\omega)$ range from -1 to 1. In general,

- $\hat{\kappa}(\omega) > 0$ indicates better than chance agreement;
- $\hat{\kappa}(\omega) = 0$ indicates exactly chance agreement;
- $\hat{\kappa}(\omega) < 0$ indicates poorer than chance agreement;
- $\hat{\kappa}(\omega) = 1$ indicates perfect agreement.

Correspondences have been established between weighted kappa and the

Pearson correlation coefficient, and between weighted kappa and the intraclass correlation coefficient (ICC). Cohen (1968) has shown that, assuming the marginal distributions are the same (i.e., $p_{i.} = p_{.i}$ for $i = 1, 2, \cdots, L$) and using the set of weights

$$\omega_{ij} = 1 - \left(\frac{i-j}{L-1}\right)^2,$$

the weighted kappa is precisely equal to the Pearson correlation coefficient calculated on integer-valued categories. And for these same weights ω_{ij}, Fleiss and Cohen (1973) have shown that, under a random-effect model, the estimate of ICC differs from $\hat{\kappa}(\omega)$ by a term involving the factor $\frac{1}{n}$ and hence is asymptotically equal to $\hat{\kappa}(\omega)$.

Thus, weighted kappa is equivalent to correlation and ICC, and hence does not relieve us from the problems noted in previous sections.

2.4. REGRESSION

Linear regression analysis, which is another commonly used approach in comparison study, should be used with caution.

Note that comparison of paired measurements in the present context is very different from the calibration problem, in which a set of measurements

are compared with and adjusted to the known true measurements, made by a standard, precise method. Misunderstanding the desirable question would lead to an inappropriate analysis.

If measurement (X) were free of error, we might fit for given data a "best" line $Y = \alpha + \beta X$, using least-squares regression. Then we might argue that this regression line should go through the origin and have a slope of one, unless there is some systematic bias. Hence we might interpret the intercept and slope — specifically, the quantities $\alpha - 0$ and $\beta - 1$, respectively — as the constant error (or relative bias) and "proportional error" [Cassidy, Triplett and LaDuca (1985), and Rawles (1986)].

However, since both sets of measurements are subject to error, necessarily $E(\beta) < 1$ and $E(\alpha) > 0$ [Altman and Bland (1983)]. Thus, the usual regression analysis would give misleading results: both relative bias and "proportional error" are expected to be non-zero, no matter how well the two sets of measurements agree.

Techniques have been developed for computing a consistent estimate of the slope of the line relating two variables, when both are subject to errors. In particular, distinct methods were developed by Bartlett (1949), Deming

(1943), and Mandel (1964); also see survey papers by Madansky (1959) and Mandel (1984) and bibliography for Chapter 1 of Draper and Smith (1981). Once a slope estimate $\hat{\beta}$ is obtained, we can estimate the intercept by $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$. But these approaches and ordinary least-squares regression all assume that measurement errors

i. follow a Guassian distribution, and

ii. are identically distributed, regardless of the sizes of items measured.

These two assumptions (especially the latter) in general do not hold in the present context. For instance, if we consider two counting processes, it is unlikely to have a discrepancy as great as 50 items for a shipment of size 100, but would be more likely for a shipment of size 1000. Thus, the variance of discrepancy for two counting processes would likely depend on the sizes of items measured; and the X, Y scatter plot would be heteroscedastic ("hetero" means "different", "scedastic" means "scatter" [Freedman, Pisani and Purves (1978, p.178)]).

2.5. WEIGHTED LEAST-SQUARES ANALYSIS

If there were no error in the X measurements, then weighted least-

squares regression would be appropriate for heteroscedastic data. In the weighted least-squares analysis to fit $Y = \alpha + \beta X$, the sum of squares to be minimized is

$$\sum_{i=1}^{n} \omega_i (Y_i - \alpha - \beta X_i)^2,$$

where usually $\omega_i = \frac{1}{\sigma_i^2}$.

If the set of weights $\{\omega_i,\ i = 1, 2, \cdots, n\}$ were known, then the solution would be

$$\hat{\alpha}_\omega = \bar{Y}_\omega - \hat{\beta}_\omega \bar{X}_\omega,$$

$$\hat{\beta}_\omega = \frac{\sum_{i=1}^{n} \omega_i (X_i - \bar{X}_\omega)(Y_i - \bar{Y}_\omega)}{\sum_{i=1}^{n} \omega_i (X_i - \bar{X}_\omega)^2},$$

where $\bar{X}_\omega = \frac{\sum_{i=1}^{n} \omega_i X_i}{\sum_{i=1}^{n} \omega_i}$, and $\bar{Y}_\omega = \frac{\sum_{i=1}^{n} \omega_i Y_i}{\sum_{i=1}^{n} \omega_i}$.

But the variance σ_i^2 and hence ω_i usually are unknown. Estimation would require iteration, using $\omega_i = (\alpha + \beta x_i)^{-1}$ or $\omega_i = (\alpha + \beta x_i)^{-2}$, etc. Notable works include Jacquez, Mather and Crawford (1968), Bement and Williams (1969), and Amemiya (1973).

But weighted least-squares, like ordinary least-squares regression, still may be inappropriate if the X_i are subject to error.

2.6. ANALYSIS OF VARIANCE AND GENERAL LINEAR MODEL

Suppose W_{ij} denotes measurement of object i made by method j, ($i = 1, 2, \cdots, n$, and $j = 1, 2$), and μ_i denotes the true but unknown value for the object i. Then the general linear model relating W_{ij} to μ_i is

$$W_{ij} = \alpha_j + \beta_j \mu_i + \varepsilon_{ij},$$

where α_j and β_j are parameters that jointly describe the measurement bias for method j, and where ε_{ij} is a random error in measuring object i with method j. It is assumed throughout that $\varepsilon_{ij} \overset{\text{indep}}{\sim} \mathcal{N}(0, \sigma_j^2)$.

This model includes the previous discrepancy model as a special case with

$$X_i = W_{i1} = \alpha_1 + \beta_{i1} + \varepsilon_{i1},$$

$$Y_i = W_{i2} = \alpha_2 + \beta_{i2} + \varepsilon_{i2}, \text{ and}$$

$$D_i = X_i - Y_i = W_{i1} - W_{i2}$$

$$= (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)\mu_i + (\varepsilon_{i1} - \varepsilon_{i2}).$$

- The model is said to have common precision if σ_j^2 is the same for all j.

- The measuring method j is said to be unbiased if $\alpha_j = 0$, and $\beta_j = 1$.

- The measuring method j is said to have a constant bias if $\alpha_j \neq 0$, but $\beta_j = 1$.

- Two methods j and k are said to have a constant relative bias if $\alpha_j \neq \alpha_k$, but $\beta_j = 1 = \beta_k$.

- The method j is said to have a nonconstant bias if $\beta_j \neq 1$.

- Two methods j and k are said to have nonconstant relative bias if $\beta_j \neq \beta_k$.

Two cases have been studied:

i. a fixed-effect model where μ_i are not randomly selected, and

ii. a random-effect model where μ_i are randomly selected from some Guassian population.

Then various linear model techniques can be employed to estimate relative bias (called "contrast") $\alpha_1 - \alpha_2$. Notable works include Grubbs (1948) and Thompson (1963). But most linear model methods assume common precision, σ_j^2 the same for all j.

More importantly, in the usual linear model, the distribution of error ε_{ij} does not depend on μ_i.

2.7. PAIRWISE DIFFERENCE AND GRAPHICAL APPROACH

Altman and Bland (1983) criticized various techniques used in reliability literature and also argued that many of these studies should be more interested in relative bias. Noting that the usual X, Y scatter plot is more relevant to correlation than to study of relative bias between the paired measurements, Altman and Bland relied on the "average-difference plot", which is a graph of pairwise difference (or discrepancy) against the average of the pair (estimate of the true measurement). One advantage of this plot is that it exhibits any trend relating discrepancy and size of measurement in a clear manner. Similar plots have been used by other statisticians, as discussed in the next chapter.

Further, Altman and Bland (1983) suggested using Pearson correlation coefficient between discrepancy $(Y - X)$ and average $\frac{X+Y}{2}$ (or sum $X+Y$) to check for equality of the total variance of the two sets of measurements. This is based on the following results. If $Var(X_i) = Var(Y_i)$, then $Cov[(X_i + Y_i), (Y_i - X_i)] = 0$; or equivalently, if $Cov[(X_i + Y_i), (Y_i - X_i)] \neq 0$, then

$Var(X_i) \neq Var(Y_i)$.

When there is no clear relationship between discrepancy and average, Altman and Bland (1983 and 1986) suggested using the normal percentile to construct a 95% confidence interval for relative bias: $(\bar{D} - 1.96S, \bar{D} + 1.96S)$, where

$$\bar{D} = \bar{X} - \bar{Y}, \text{ and}$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2.$$

Essentially, the central limit theorem is applied here. In the following chapter, we consider also nonparametric confidence intervals for median discrepancy.

Altman and Bland (1983) proposed using transformation of the data if the "average-difference plot" indicates any relationship between the discrepancy and the average. However, no example has been shown. Indeed, an appropriate transformation may not be obvious. Also, if the discrepancies are symmetrically distributed, transformation that destroys symmetry may not be desirable.

2.8. QUESTIONS BEYOND ALTMAN AND BLAND

There is a serious issue not much dealt with even by Altman and Bland.

For paired measurements X, Y, scatter plot often shows heteroscedasticity — plots described in Chapters 5 and 7 demonstrate the issue dramatically. Some reliability techniques allow for heteroscedasticity (using weighted least-squares regression, for example); but the normal confidence interval above does not.

More precisely, consider paired measurements X_i and Y_i, and the corresponding discrepancy D_i, such that the variance

$$Var(D_i) = Var(X_i) + Var(Y_i) - 2Cov(X_i, Y_i)$$

is a function of the true magnitude μ_i; then how should we interpret a "sample" variance of $D_1, D_2, \cdots, D_n$, or a "sample" correlation for $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, if the magnitudes $\mu_1, \mu_2, \cdots, \mu_n$ have been arbitrarily or intentionally (but not randomly) selected?

This thesis, using a perspective related to permutation tests, tries to estimate the relative bias without any assumption about the mechanism by which the items measured are chosen.

3. INDEPENDENT, IDENTICALLY DISTRIBUTED DISCREPANCIES: A SIMPLE MODEL

Suppose $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ are the pairwise observations obtained by two raters or two measuring processes on n objects. Then their discrepancies $D_1, D_2, \cdots, D_n$ are given by

$$D_i = X_i - Y_i, \qquad i = 1, 2, \cdots, n.$$

As discussed earlier, agreement or disagreement may be characterized primarily by a central location parameter such as the mean (that is, relative bias) or the median of the discrepancy distribution. In various contexts, we may wish to estimate this location parameter, say θ; or to provide a confidence interval; or to test a hypothetical value of this parameter (usually $\theta = 0$ would be of interest).

Our statistical concern here is comparison of observers or of measurement methods — not the "true" magnitudes (say $\mu_1, \mu_2, \cdots, \mu_n$) of the particular objects being measured, nor the separate distributions of X_i, Y_i measurements (with, say, $X_i = \mu_i + \varepsilon_i$ and $Y_i = \mu_i + \delta_i$ for some errors ε_i, δ_i).

But, in general, the underlying distribution of discrepancy D_i could

depend on the magnitude μ_i being measured. For instance, the standard deviation of the discrepancy distribution may be proportional to the magnitude of the object being measured or to the square root of that magnitude, etc. The simplest model, which is the subject of this chapter, would assume that the discrepancy distribution is *not* a function of the quantity being measured. This assumption, together with the independence assumption, models $D_1, D_2, \cdots, D_n$ as independent and identically distributed observations. This assumption holds, in particular, if the (X_i, Y_i) are independent and identically distributed random vectors.

Even if the observed discrepancies $D_1, D_2, \cdots, D_n$ are independent and identically distributed, the underlying distribution, in general, is still unknown and may be in any shape. But an estimator, confidence interval or hypothesis test for the central location may be more or less efficient, relative to some other method, depending on whether the unknown distribution is symmetric or skewed, whether it is light-tailed or heavy-tailed, and so on. In this chapter, we consider alternative (or competitive) estimators, etc.

3.1. MEAN AND t PROCEDURES

The expected value or mean of a distribution is the parameter most

often used to characterize central tendency. And the usual unbiased estimate of the distribution mean is the sample mean. Given a sample of discrepancies, $D_1, D_2, \cdots, D_n$, the sample mean, $\bar{D}$, is defined to be

$$\begin{aligned} \bar{D} &= \text{mean of } \{D_1, D_2, \cdots, D_n\} \\ &= \frac{1}{n} \sum_{i=1}^{n} D_i \ . \end{aligned}$$

We would also like to construct a confidence interval (or interval estimate) for the distribution mean; hence, we need the distribution as well as the expected value of the sample mean $\bar{D}$. If the discrepancies are normally distributed, then the sample mean, $\bar{D}$, also will be normally distributed. However, since we need to use the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2$$

to estimate the unknown variance of the normal discrepancies, a confidence interval for the location parameter of the discrepancy distribution is obtained based on the Student t distribution, with $n-1$ degrees of freedom. Hence, a $1-2\alpha$ symmetric confidence interval, $(\theta_{LOW}, \theta_{UP})$, is given by

$$\theta_{LOW} = \bar{D} - t_{\alpha,n-1} \frac{S}{\sqrt{n}} \quad \text{and} \quad \theta_{UP} = \bar{D} + t_{\alpha,n-1} \frac{S}{\sqrt{n}},$$

where $t_{\alpha,n-1}$ is the upper 100α percentile point (or denotes the $100(1-\alpha)$ ordinary percentile) of the t distribution, with degrees of freedom $n-1$.

For large n, percentile $t_{\alpha,n-1}$ can be approximated by z_α, the corresponding percentile of the standard normal distribution. (In particular, $z_\alpha = 1.96$ if $\alpha = 0.025$.)

To test the hypothesis

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0,$$

at 2α level of significance, we construct the t statistic

$$t = \frac{\bar{D} - \theta_0}{S/\sqrt{n}}$$

and compare it with $t_{\alpha,n-1}$. We would reject H_0 and conclude that θ is significantly different from θ_0 at the 2α level of significance if and only if $|t| > t_{\alpha,n-1}$.

If the underlying discrepancy distribution is not normal, but is at least symmetric, then the t test can be regarded as a permutation test, discussed below, although the nominal significance level would not be exact.

Even without symmetry, the central limit theorem still would provide approximate normality for the sampling distribution of $\bar{D}$, assuming only that the discrepancy distribution has finite variance. Thus, asymptotically, the situation would be the same as the previous case, and the confidence interval and hypothesis testing can be based on the t distribution as before.

There is considerable literature on how non-normality affects the t statistic and confidence intervals. Notable are works by E. S. Pearson and Adyanthaya (1929), Geary (1936, 1947), Gayen (1949), Efron (1969), E. S. Pearson and Please (1975), and Cressie (1980). This literature indicates that asymmetry (or skewness) of the underlying distribution affects the distribution of t more than does the kurtosis (heavy- or light-tailedness).

In the present context, the underlying distribution characterizes difference between two measuring processes, X and Y. If measurements X and Y have distributions of the same shape, but shifted — that is, if the processes have different biases, but the same variance — then the distribution of difference, $X - Y$, must be symmetric [Pratt and Gibbons (1981, p.147)]. Thus, the t confidence interval and the t test for the discrepancies tend to be robust for inference with respect to measurement discrepancies.

3.2. MEDIAN AND SIGN PROCEDURES

The median is another well known parameter characterizing the central location of a distribution. By definition, the median of the distribution of D is a point d such that

$$Prob(D < d) \leq \frac{1}{2} \leq Prob(D \leq d).$$

Notice that the median, in general, is not uniquely defined. However, if the underlying distribution is continuous, then the median is unique and can be defined as that value d such that

$$Prob(D \leq d) \;=\; \frac{1}{2}.$$

For symmetric distribution, median and mean are the same value, the symmetry point (provided that the expectation exists and is finite).

One simple estimator of the distribution median would be the sample median. Given a sample of discrepancies, $D_1, D_2, \cdots, D_n$, the sample median, $\tilde{D}$, is given by

$$\begin{aligned} \tilde{D} &= \text{median of } \{D_1, D_2, \cdots, D_n\} \\ &= \begin{cases} D_{(\frac{n+1}{2})}, & \text{if n is odd;} \\ \frac{1}{2}\left[D_{(\frac{n}{2})} + D_{(\frac{n}{2}+1)}\right], & \text{if n is even.} \end{cases} \end{aligned}$$

Here, $D_{(i)}$ denotes the i^{th} order statistic of $D_1, D_2, \cdots, D_n$.

Furthermore, a confidence interval can be obtained based on order statistics. Suppose $D_1, D_2, \cdots, D_n$ are the observed discrepancies; then a $1-2\alpha$ symmetric confidence interval $(\theta_{LOW}, \theta_{UP})$ is given by

$$\theta_{LOW} = D_{(n+1-b_\alpha)} \quad \text{and} \quad \theta_{UP} = D_{(b_\alpha)},$$

where b_α is the upper 100α percentile point of the binomial distribution

with sample size n and $p = \frac{1}{2}$. That is, b_α is the value such that

$$Prob(B \geq b_\alpha) \;=\; \alpha,$$

where $B \sim Bin\left(n, \frac{1}{2}\right)$ [Hollander and Wolfe (1973, pp.48–49)]. This binomial percentile point can be obtained from tables of the binomial distribution or of the incomplete beta function [e.g., Harvard (1955) or Owen (1962)].

For large n, the integer b_α can be approximated by

$$b_\alpha \;\approx\; \frac{n}{2} + 1 + z_\alpha \sqrt{\frac{n}{4}}\ ,$$

where z_α is the standard normal percentile (defined before). The value on the right hand side, in general, is not an integer, so in practice the closest integer is used. This gives a large-sample approximate confidence interval for the median discrepancy [Hollander and Wolfe (1973, p.49)].

If the problem of interest is to test the hypothesis

$$H_0 : \theta = \theta_0 \;\text{ versus }\; H_1 : \theta \neq \theta_0,$$

then we define the sign statistic

$$B \;=\; \sum_{i=1}^{n} I(D_i > \theta_0),$$

where the indicator function

$$I(D_i > \theta_0) = \begin{cases} 1, & \text{if } D_i > \theta_0; \\ 0, & \text{if } D_i < \theta_0, \end{cases}$$

and reject H_0 (to conclude that θ is significantly different from θ_0) at the 2α level of significance if either $B \geq b_\alpha$ or $B \leq n - b_\alpha$ [Hollander and Wolfe (1973, p.40)].

In computing the sign statistic B, above, $I(D_i > \theta_0)$ has not been defined when $D_i = \theta_0$. We can avoid this difficulty, in theory, by assuming continuous distribution. In practice, measurements are not always sufficiently precise to avoid zeros, even if the distribution is continuous. For methods of handling zeros in the sign test, see Hemelrijk (1952), Putter (1955), Noether (1967), Krauth (1973), and Pratt and Gibbons (1981, pp.97–104).

3.3. THE HODGES-LEHMANN ESTIMATOR AND SIGNED RANK PROCEDURES

The procedures discussed below compromise between the t and sign procedures: the underlying distribution should be symmetric (or approximately symmetric), but normality is not needed. And, as noted in Section 3.1, the discrepancies would be symmetrically distributed if the distributions of X and Y have the same shape and differ only by a location shift.

For a sequence of observations $D_1, D_2, \cdots, D_n$, define the set of Walsh averages, the $m = \dfrac{n(n+1)}{2}$ quantities

$$\left\{ \frac{D_i + D_j}{2}, \;\; i, j = 1, 2, \cdots, n \;\; \& \;\; i \leq j \right\}.$$

Then the corresponding Hodges-Lehmann statistic $\hat{D}$ is defined to be the sample median of these Walsh averages, that is

$$\hat{D} \;\; = \;\; \text{median of} \;\; \left\{ \frac{D_i + D_j}{2}, \;\; i, j = 1, 2, \cdots, n \;\; \& \;\; i \leq j \right\}.$$

If the underlying discrepancy distribution is symmetric, then the Hodges-Lehmann statistic (the Walsh median) estimates this centre (= median = mean) [Hollander and Wolfe (1973, p.33)].

Moreover, a symmetric confidence interval for the symmetry point can be based on the Walsh averages: for confidence level $1 - 2\alpha$ the interval $(\theta_{LOW}, \theta_{UP})$ is given by

$$\theta_{LOW} = W_{(m+1-w_\alpha)} \quad \text{and} \quad \theta_{UP} = W_{(w_\alpha)},$$

where $W_{(1)} \leq W_{(2)} \leq \cdots \leq W_{(m)}$ denote the ordered Walsh averages, with $m = \dfrac{n(n+1)}{2}$, and w_α is the upper 100α percentile point of the Wilcoxon signed rank statistic, whose exact distribution is available in tables [e.g., Owen (1962) or Pearson and Hartley (1972)]. For large n, the integer w_α

can be approximated by

$$w_\alpha \approx \frac{n(n+1)}{4} + 1 + z_\alpha \sqrt{\frac{n(n+1)(2n+1)}{24}},$$

where z_α is the standard normal percentile and the right hand side is rounded to the closest integer value [Hollander and Wolfe (1973, pp.35–36)].

To test the hypothesis

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0,$$

we define the Wilcoxon signed rank statistic

$$W = \sum_{i=1}^{n} R_i I(D_i > \theta_0),$$

where R_i denotes the rank of $|D_i|$ in the ranking from least to greatest of absolute values $|D_1|, |D_2|, \cdots, |D_n|$, and the indicator function

$$I(D_i > \theta_0) = \begin{cases} 1, & \text{if } D_i > \theta_0; \\ 0, & \text{if } D_i < \theta_0. \end{cases}$$

We would reject H_0 (to conclude that θ is significantly different from θ_0) at the 2α level of significance if either $W \geq w_\alpha$ or $W \leq m - w_\alpha$ [Hollander and Wolfe (1973, p.28)].

Zero values may be a practical problem for the signed rank procedure (as for the sign test). Also non-zero ties (two or more observations which

have the same magnitude) can cause complications for Wilcoxon signed rank procedures. For present purposes, "midranks" as defined by Lehmann (1975) may be used when ties render "rank" ambiguous. For discussion of zeros and ties, see Conover (1973), Cureton (1967), Pratt (1959), and Rahe (1974).

3.4. DISCUSSION: THE PERMUTATION PERSPECTIVE

Obviously, normality is the most restrictive assumption considered above; and the ordinary median estimator and the sign test are the least restricted procedures, not even requiring symmetry. The signed rank procedures are intermediate.

If the underlying distribution is symmetric, then all three approaches considered above — that is: the t test, sign test, and signed rank test — are permutation (or randomization) procedures, corresponding to different score functions. For a random vector $D = (D_1, D_2, \cdots, D_n)$, a permutation test statistic or "generalized Student's statistic" [Efron (1969)] has the form

$$\tilde{S}_n = \sum_{i=1}^{n} \tilde{U}_i,$$

where the vector

$$\tilde{U} = g(U), \quad \text{and}$$

$$U = \frac{D}{\|D\|} = \frac{D}{\sqrt{\sum_{i=1}^{n} D_i^2}},$$

and g is a symmetry preserving transformation of the unit n-sphere into itself. For instance, if

$$g^+(\xi) = \sqrt{\frac{2}{n(n+1)}}\ (R_1, R_2, \cdots, R_n)$$

is defined on the positive orthant $S_n^+ = \{(\xi_1, \xi_2, \cdots, \xi_n) : \xi_i > 0, \sum_{i=1}^n \xi_i^2 = 1\}$, where R_i is the rank of ξ_i among $\{\xi_1, \xi_2, \cdots, \xi_n\}$ and g maps every orthant into itself in a similar fashion, then $\tilde{S}_n$ is the Wilcoxon signed rank statistic.

The permutation perspective is important since, as pointed out in Section 3.1, the discrepancy distribution will be symmetric if the X and Y distributions have the same shape, but possibly differ by a location shift.

Choice among different permutation procedures should be based on relative performance as discussed in the next chapter.

3.5. BOOTSTRAP METHOD

Efron's "bootstrap", related to Tukey's "jackknife", provides a general method to construct nonparametric estimators or confidence intervals [Efron (1979)]. The bootstrap method can be applied to estimate any parameter characterizing central tendency; but the bootstrap estimators of distribution mean and median turn out to be essentially the same as the ordinary sample

mean and sample median. Hence, the general theory of the bootstrap is not utilized in the remainder of this thesis. (But discussion of the bootstrap method is provided in Appendix I.)

The bootstrap has been modified to utilize partial knowledge about the underlying distribution of interest: but the "smoothed bootstrap" [Efron (1981)], the "parametric bootstrap" [Efron (1985)] and the "Bayesian bootstrap" [Rubin (1981)], will not be discussed here.

4. EVALUATING RELATIVE PERFORMANCE OF CONFIDENCE INTERVAL PROCEDURES

Choice among competing statistical procedures should be based on their relative performance. In statistical literature, relative efficiency is most often defined in the hypothesis testing context. This chapter first reviews relative efficiency of tests and an equivalent definition of relative efficiency based on expected ratio of squared lengths of confidence intervals. Also, we propose two other criteria of relative performance: probability that one procedure produces a shorter confidence interval than the other procedure, and the relative variability of confidence interval lengths. (These criteria are easier to interpret than other relative efficiency definitions, such as by Bahadur, Hodges-Lehmann, or Chernoff, etc.)

4.1. RELATIVE EFFICIENCY OF STATISTICAL TESTS

Relative efficiency of two statistical tests (e.g., sign test and signed rank test) in general would depend on

i. the specified significance level,

ii. the alternative hypothesis value,

iii. the sample size, and

iv. the form of the underlying distribution.

Pitman (1948) defined an asymptotic relative efficiency (ARE) which depends only on the form of the underlying distribution, but requires symmetry. Essentially, Pitman efficiency of test procedure T_2 with respect to T_1, denoted here by $ARE(2,1)$, is equivalent to the limiting ratio of sample sizes, $\frac{n_1}{n_2}$, such that both tests achieve equal power against a sequence of alternatives that are "close" to and approaching the null hypothesis [Randles and Wolfe (1979, pp.142–144)]. Note that, if $ARE(2,1)$ is the efficiency of T_2 relative to T_1, then $ARE(1,2) = \frac{1}{ARE(2,1)}$. Pitman efficiency can be represented as a ratio of *efficacies*, defined in Appendix II and evaluated in Table 1 for permutation tests applied to familiar distributions.

Based on Pitman efficiency, statistical literature (notably literature on nonparametric methods) gives the following general recommendations for the t test (T), sign test (S), and Wilcoxon signed rank test (W) [Randles and Wolfe (1979, pp.166–168)].

- T is optimal for normal distribution and performs well for other distributions with moderate tails (e.g., logistic distribution);

- T is preferable to S and is comparable to W for distributions with light tails (e.g., uniform distribution);

- T is inferior to both S and W for distributions with heavy tails (e.g., Cauchy or double exponential distribution);

- S is preferable for distributions with very heavy tails;

- W is intermediate between S and T, and therefore is "robust" in the sense of being a good compromise.

However, there are few guidelines for a mixture of normals or for a contaminated normal distribution, which will be of interest in the next chapter.

4.2. RELATIVE EFFICIENCY OF CONFIDENCE INTERVAL PROCEDURES

In estimation or data analysis context, Pitman ARE can be interpreted in terms of lengths of confidence intervals. Suppose $L_{1,n}$ and $L_{2,n}$ are the respective lengths of confidence intervals for θ, corresponding to tests T_1 and T_2, respectively, and both based on the same sample of size n. If T_1

produces a confidence interval with expected length less than that produced by T_2, then we say T_1 is more efficient than T_2. It can be shown that, under suitable conditions, $\left(\frac{L_{1,n}}{L_{2,n}}\right)^2$ converges to $ARE(2,1)$ in probability, as $n \to \infty$ [Pratt and Gibbons (1981, p.376)]. It follows that

$$E_\theta\left[\left(\frac{L_{1,n}}{L_{2,n}}\right)^2\right] \longrightarrow ARE(2,1) \ ,$$

or equivalently

$$E_\theta\left(\frac{L_{1,n}}{L_{2,n}}\right) \longrightarrow \sqrt{ARE(2,1)} \ .$$

In confidence interval context, it seems natural to consider lengths rather than squared lengths. Thus, the asymptotic expectation of $\frac{L_1}{L_2}$ (suppressing notational dependence on n) is an important criterion of relative performance.

Pratt (1961) provided another interpretation for ARE: relative probability of including a false value. Pratt showed that

$$\begin{aligned} E_{\theta_0}(\theta_{UP} - \theta_{LOW}) &= \int_\theta P_{\theta_0}(\theta_{LOW} \le \theta \le \theta_{UP})\, d\theta \\ &= \int_{\theta \neq \theta_0} P_{\theta_0}(\theta_{LOW} \le \theta \le \theta_{UP})\, d\theta, \end{aligned}$$

where $(\theta_{LOW}, \theta_{UP})$ is a confidence interval for θ, and θ_0 is the true (but unknown) value of θ. Notice that the last integral gives the probability of including a particular false value and "averages" over all possible false values [Pratt and Gibbons (1981, p.50)].

Recall that Pitman ARE provides an asymptotic comparison. How applicable are these asymptotic results for the finite sample size n? Monte Carlo results (see below) show that, in general, the finite-sample behaviour is reasonably close to asymptotic results; but anomalies may arise.

4.3. OTHER CRITERIA OF RELATIVE PERFORMANCE

Instead of comparing lengths in expectation, we can compare them in probability: that is, consider the probability that procedure T_1 produces a confidence interval shorter than that produced by T_2. If $L_1 < L_2$ with probability much greater than 0.5, i.e., if

$$Prob_\theta\,(L_1 < L_2) \gg 0.5,$$

then procedure T_1 may be preferred to T_2 even if their ARE is close to 1.

Besides considering which confidence interval is shorter in expectation or in probability, we would also like to have an interval whose length has relatively small variance. Thus, relative variability (or inversely, stability) of confidence interval lengths provides another criterion of performance. If the standard deviation of L_1 is much less than that of L_2, i.e., if

$$\frac{SD(L_1)}{SD(L_2)} \ll 1,$$

we would conclude that T_1 performs better than T_2.

Unluckily, exact theoretical results for the probability and relative variability criteria are very difficult, if not impossible, to obtain. For instance, if the confidence intervals related to the sign test and to the Wilcoxon signed rank test are compared, difficult integrals based on certain joint distributions of order statistics are required [Sarhan and Greenberg (1962)]. In order to study these three criteria for specific distributions, Monte Carlo simulations are needed to approximate the theoretical exact results.

In summary, we would prefer T_1 to T_2 if

i. $E\left(\frac{L_1}{L_2}\right) \ll 1$,

ii. $P(L_1 < L_2) \gg 0.5$, and

iii. $\frac{SD(L_1)}{SD(L_2)} \ll 1$.

But, in Monte Carlo studies below, there are examples in which one criterion favours T_1 while another favours T_2. Also, there are examples of distributions for which

- $E(L_1/L_2) \approx 1$ and $SD(L_1)/SD(L_2) \approx 1$, but $P(L_1 < L_2) \gg 0.5$, or

- $E(L_1/L_2) \approx 1$ and $P(L_1 < L_2) \approx 0.5$, but $SD(L_1)/SD(L_2) \ll 1$.

So it is clear that the three criteria, above, do not imply one another; and, in particular, both of our new criteria may have practical importance: to choose between two procedures when relative efficiency is approximately equal to 1.

4.4. EVALUATION OF THE THREE PERFORMANCE CRITERIA FOR SPECIAL DISTRIBUTIONS: MONTE CARLO RESULTS

In order to examine the relevance of asymptotic results and to evaluate the above criteria for specific distributions, we consider the following Monte Carlo studies. One thousand random samples, each of size $n = 32$, are generated from each of eleven distributions: standard normal $\mathcal{N}(0,1)$, uniform$(-1,1)$, Cauchy (or t with one degree of freedom), equal-proportion mixtures of four and five normals, where each component is $\mathcal{N}(0,i^2)$ with $i = 1,2,3,4$ and $i = 1,2,3,4,5$, and six contaminated standard normals. (Mixed and contaminated distributions are useful to approximate various situations in which the assumption of "identically distributed observations" is

not valid. Efficiency results for mixtures also demonstrate importance of the new performance criteria.) Notice that the normal, uniform, and Cauchy are examples of distributions with medium, light, and heavy tails, respectively. Since the three criteria of relative performance do not depend on the distribution parameters (except for the mixtures of normals and contaminated normals), there is no loss of generality in considering *standard* normal, uniform, and Cauchy. For interval estimation corresponding to the exact sign test (not using normal approximation), exact 95% coverage in general is not attainable (without randomization); but for sample size $n = 32$, coverage of 94.98% is an attainable level. Also, $n = 32$ is close to sizes of some real data sets considered below.

For the distributions just described, we consider confidence intervals corresponding to the t test (T), the sign test (S) and the Wilcoxon signed rank test (W). All three confidence interval procedures are available in the Minitab (version 5) statistics package [Ryan, Joiner and Ryan (1985)], but for present purposes have been programmed in the "S" statistics environment under UNIX [Becker and Chambers (1984)]. For these three confidence interval procedures, asymptotic relative efficiency (ARE) and simulation results (for $n = 32$) are shown in Tables 2 - 6. (Note that Table 2 gives $ARE(W, S)$ while $ARE(S, W) = \dfrac{1}{ARE(W, S)}$ is the $W : S$ squared length ra-

tio, given by Table 3; etc.). Table 3 gives length and squared length ratios for finite samples ($n = 32$), results directly comparable to the asymptotic ratios in Table 2. Tables 4 and 5 compare intervals using variability and probability criteria for finite samples. Table 5 also gives actual coverages of the nominal 95% confidence intervals (each entry based on 1000 samples, with sample size $n = 32$).

4.4.1. COMPARISON OF ASYMPTOTIC AND FINITE-SAMPLE RESULTS

- In the Monte Carlo experiment with 1000 replications, most of the actual coverages are within 1% of the nominal coverage, 95%, except for the T interval with Cauchy distribution. Observed coverage of T for Cauchy distribution (98%) confirms that T is conservative for a long-tailed distribution [Benjamini (1983)].

- In general, if one procedure is preferred to another by Pitman's asymptotic relative efficiency criterion, then the same preference holds for $n = 32$. But the "advantage" may be consistently and considerably less (or more) for finite samples than asymptotically. For instance: "advantage" of W over S is less for $n = 32$ than

asymptotically; "advantage" of T over S is less for $n = 32$ than asymptotically; and "advantage" of W over T is less for $n = 32$ than asymptotically, for most distributions.

- Reversals are technically possible, an example being the mixture of four normals: asymptotically W is more efficient than S (the length ratio $L_W/L_S \to \sqrt{0.9693} < 1$ as $n \to \infty$); but for $n = 32$, the Monte Carlo sample mean of $(L_W/L_S) = 1.0402$.

4.4.2 EXAMPLES: NEW CRITERIA MAY BE DECISIVE

Recall that, in addition to the usual definition of relative efficiency as expected ratio of squared interval lengths, we wish to consider the probability that one procedure produces an interval shorter than the other and the relative variability of interval lengths. In the simulation results, we evaluate not only the sample mean of the ratio of lengths, but also the percentage of times such that one interval is shorter than the other, and the ratio of sample standard deviations of lengths.

The following examples call attention to criteria other than usual relative efficiency.

EXAMPLE 1A: Probability criterion may be critical.

Consider W and T for uniform distribution. The Monte Carlo experiment with 1000 replications gives

$$E\left(\frac{L_W}{L_T}\right) \approx 1.09 \approx 1, \text{ and } \frac{SD(L_W)}{SD(L_T)} \approx 1.19 \approx 1,$$

which do not suggest strong preference for T. But since $L_T < L_W$ for 968 out of 1000 simulated samples, T would be preferred by the probability criterion.

EXAMPLE 1B: Probability criterion may be critical.

Consider W and T for the mixture of five normals. The Monte Carlo experiment with 1000 replications gives

$$E\left(\frac{L_W}{L_T}\right) \approx 0.92 \approx 1, \text{ and } \frac{SD(L_W)}{SD(L_T)} \approx 1.01 \approx 1,$$

which do not suggest strong preference for W. But since $L_T < L_W$ only for 232 out of 1000 samples, W would be preferred by this criterion.

These examples suggest that the probability criterion is more sensitive than the expectation criterion (usual relative efficiency) with respect to the shape of the distribution tails.

EXAMPLE 2A: Variability criterion may be critical.

If W and S are compared for the mixture of five (or four) normals, simulation results give

$$E\left(\frac{L_W}{L_S}\right) \approx 1.06 \text{ (or } 1.04) \approx 1, \text{ and}$$

$$L_W < L_S \text{ for } 470 \text{ (or } 518), \approx 50\% \text{ of } 1000 \text{ samples.}$$

But since, for both mixtures,

$$\frac{SD(L_W)}{SD(L_S)} \approx 0.65 \ll 1,$$

the relative variability of interval lengths clearly favours W, although the other two criteria do not provide any clear preference.

EXAMPLE 2B: Variability criterion may be critical.

For the mixture of five normals

$$E\left(\frac{L_T}{L_S}\right) \approx 1.18, \text{ and}$$

$$L_T < L_S \text{ for only } 337 \text{ out of } 1000 \text{ samples,}$$

so these two criteria slightly favour S. But

$$\frac{SD(L_T)}{SD(L_S)} \approx 0.64 \ll 1$$

and supports use of T. We might accept slightly greater length in order to reduce variability of our confidence interval.

EXAMPLE 2C: Variability criterion may be critical — performance of W versus T for contaminated normals.

As discussed in Section 4.1, based on Pitman efficiency, T is slightly preferable to W for [pure] normal distributions, with $ARE(W,T) = 0.95$ or, equivalently, $\frac{L_W}{L_T} \to 1.02$; and simulation results for $n = 32$ give

$$E\left(\frac{L_W}{L_T}\right) \approx 1.03,$$

$$L_T < L_W \text{ for } 731 \text{ out of } 1000 \text{ samples, and}$$

$$\frac{SD(L_W)}{SD(L_T)} \approx 1.13.$$

However, normality is a strong assumption — in practice, a small percentage of contaminant is common (e.g., a small percentage of measurements with large errors; or a small percentage of time when a process is "out of control").

For example, consider simulation results for the *standard* normal *contaminated* with 5% of $\mathcal{N}(0, 16)$:

$$E\left(\frac{L_W}{L_T}\right) \approx 0.91, \text{ and}$$

$$L_T < L_W \text{ for } 393 \text{ out of } 1000 \text{ samples,}$$

so these two criteria slightly favour W. But since

$$\frac{SD(L_W)}{SD(L_T)} \approx 0.49 \ll 1$$

the relative variability of interval lengths strongly favours W.

4.5. DISCUSSION

Pitman ARE, which is equivalent to the limiting ratio of squared lengths of two confidence interval procedures, is not the only measure of relative performance. Fortunately, in our examples, the asymptotic efficiency generally is similar to the corresponding expected ratio for finite sample size. But, in addition to expectation of the length ratio, we may wish to consider a) the probability that this length ratio exceeds one, and b) the relative variability of lengths of the two confidence intervals. In Monte Carlo simulation, these three criteria of relative performance generally complement each other, but there are instances which spotlight the differences.

Overall, signed rank confidence interval W is generally good: even in the worst case, the longer intervals produced by W have expected length not much longer nor much more variable than the competitive procedures.

4.6. EPILOGUE: ADAPTIVE PROCEDURES

The foregoing discussion is from a traditional perspective in which the statistician's choosing of a statistical procedure (from among T, S, and W confidence interval procedures, or from any other collection) is separated from application and computation. But there have been several suggestions that would formally unify these aspects of statistical analysis. In principle, we could

- calculate several confidence intervals (e.g., T, S, and W intervals) for a given data set; and then

- calculate some auxiliary statistic(s) and invoke a formal decision rule based on such statistic to select and report one of the available intervals.

The simplest such suggestions have been the following.

- Adaptive procedures of Randles and Hogg (1973): if one method is generally good for data from light-tailed distributions and another method is good for heavy-tailed distributions, then calculate a statistic that quantifies tail weight and adopt a formal cut-off value. This idea obviously generalizes from two competitors to several. See

Appendix III.

- "Legalized cheating" proposed by Efron (1969): select the first of two available confidence intervals if and only if their length ratio $\frac{L_1}{L_2} \leq 1$. Efron suggests appropriate adjustment of confidence level, working with sign and t procedures.

These ideas are intriguing and we hope to consider them in the future — but not in the remainder of this thesis. They present both practical and theoretical problems.

Adaptive procedures introduce additional decision parameters; and also they may be computationally difficult. At present, popular statistical computation packages, such as SAS, BMDP, and SPSS do not even provide S and W confidence intervals, corresponding to the sign and Wilcoxon signed rank tests; and Minitab made these confidence intervals available only in 1985 — using normal approximation to Wilcoxon confidence interval [Ryan, Joiner and Ryan (1985, p.290)].

Little is known about relative performance criteria to choose among adaptive options. And little is known about behaviour of such procedures when data are not identically distributed, as considered in the next chapter.

5. NON-IDENTICALLY DISTRIBUTED DISCREPANCIES: A GENERAL MODEL

The simplest model assumes that discrepancy between two measuring procedures does not depend on the quantity measured, so that observed discrepancies are independent and identically distributed. This chapter considers a more general model: discrepancies that are independent, but need *not* be identically distributed. In particular, the variance of a discrepancy distribution may be a function of (for instance, proportional to) the magnitude of the object measured, so that distributions of $D_1, D_2, \cdots, D_n$ differ with respect to scale.

But, in order to have a meaningful concept of agreement to estimate or to test, the discrepancy distributions must have some location parameter in common; and this chapter will assume that the non-identical discrepancy distributions have identical means or medians. Again note that these two parameters would be equal if the discrepancies are symmetrically distributed; and that $D_i = X_i - Y_i$ will have symmetric distribution whenever the X_i and Y_i distributions have the same shape, differing only by some shift (or relative bias). Variance or other scale parameters for X_i and Y_i distributions still could be a function of the magnitude of the object measured. (In

practice, the assumption of sampling with a common median or mean often may be valid for discrepancy *rate* — difference between measurements as a percentage of quantity measured — considered in the next chapter.)

This chapter first considers empirical (largely graphical) methods to explore or validate distributional assumptions about $D_1, D_2, \cdots, D_n$, such as symmetry, functional modelling of variance, etc. A "random walk" model is then proposed to provide theoretical basis for approximate normality (and symmetry) of discrepancy distribution when X and Y are two counting processes. Assuming symmetry, choice among the three statistical procedures is discussed by comparing the non-identically distributed case to sampling from a suitable mixture. Finally, a summary guide is provided to estimate median discrepancy for real data.

5.1. GRAPHICAL METHODS FOR DISCREPANCIES

Scatter plotting of X_i and Y_i values and calculation of the usual (Pearson) correlation are commonly used to check for linear relationship and to demonstrate that one variable can predict the other. But when X and Y processes measure the same objects, strong positive correlation (i.e., high reliability) is usual — otherwise the measurements would be useless in

practice. For strongly linear X, Y points, visual resolution of discrepancy is poor. Also, the calculated correlation may be misleading: the greater the range of magnitudes measured, the better agreement will appear to be. See Altman and Bland (1983) and the discussion of correlation above, in Chapter 2.

A more effective graphical method for assessing agreement between measuring processes is to plot $Y - X$ against the average measurement $\frac{X+Y}{2}$ [Altman and Bland (1983)]. This leads to an "average-difference plot", which is a variation of the "sum-difference graph" attributed to Tukey by Cleveland (1985, pp.118–23). The average is used here in lieu of the sum in Tukey's graph, because of its obvious interpretation — as a combined estimate of the magnitude measured separately by X and Y. Either the average-difference plot or the sum-difference graph rotates a scatter plot 45° in a clockwise direction and then expands the rotated points vertically to fill the plotting region [Cleveland (1985, p.122)]. Notice that the difference $Y_i - X_i$ and the sum $X_i + Y_i$ are uncorrelated random variables if X_i and Y_i have the same variance (similarly for $X_i - Y_i$ instead $Y_i - X_i$, and for average $(X_i + Y_i)/2$ instead of $X_i + Y_i$) and hence are independent if X_i, Y_i are bivariate normal. Or, equivalently, the variances of X_i and Y_i cannot be equal if the difference $X_i - Y_i$ and the sum $X_i + Y_i$ are not uncorrelated.

It is sometimes useful also to plot absolute value $|X-Y|$ versus average or sum. For some data it may be useful to plot the abscissa on a non-linear scale, such as logarithm or square root of $X+Y$. (See the first example in Chapter 7.)

Using these plots, we can visually assess the range of measurement and check for symmetry of discrepancy; and we can see if there is any trend, for instance, whether and how magnitude of discrepancy increases with the (estimated) magnitude of the object measured.

If $D_1, D_2, \cdots, D_n$ are independent and identically distributed, then vertical scatter in the average-difference plot should be about the same over any horizontal interval, regardless of the interval's position along the abscissa. And if the data are symmetric, then scatter in the average-difference plot should be symmetric about a horizontal line. (We can also check symmetry using a q-q plot or normal probability plot of the sample distribution, as well as a histogram.)

That is, symmetric, identically distributed data tend to fill a rectangle in the average-difference plot. If instead the standard deviation of the discrepancy distribution is proportional to the magnitude of measurement,

then the average-difference plot will exhibit a "shotgun" pattern — spread out in a triangle, symmetric about a horizontal line.

Similarly, if the standard deviation of the discrepancy distribution is proportional to the square root of the magnitude measured, then the average-difference plot will diverge like a root function; or a plot of difference versus the square root of the average will fan out linearly. (This plot was used by by Professor Ned Glick for the logging data in Chapter 7.) The following section develops a corresponding theoretical model.

5.2. DISCREPANCY BETWEEN TWO *COUNTING* PROCESSES: RANDOM WALK MODEL

Suppose X and Y are integer counts of the same lot of items; for instance, in data analysis considered in Chapter 7, X and Y may be counts by two inspectors of the same batch of logs or of cells on the same laboratory slide. Then the discrepancy $X - Y$ may be considered as a sum (over items) of random differences. Such sequence is equivalent to "steps" in a "random walk" — the walk observed only once, after an unknown number of steps (corresponding to the number of items in the batch).

This model was suggested by Professor Ned Glick in analysis of the

logging data considered in Chapter 7. As shown here, this model implies that variance of discrepancy should be proportional to batch size.

Consider a batch containing B items. For each item or piece in the batch, the counting process X may miss that piece, may count it correctly once, or may double count it, etc., with some (unknown) probability distribution whose expectation is p and whose variance is r. Then the X count can be treated as the sum of these successive contributions: $X = \sum_{k=1}^{B} P_k$, where P_k is the contribution of the k^{th} piece in the lot; and

$$E(X) = E\left(\sum_{k=1}^{B} P_k\right) = \sum_{k=1}^{B} E\left(P_k\right) = Bp.$$

If the piecewise contributions P_k are independent, then

$$Var(X) = Var\left(\sum_{k=1}^{B} P_k\right) = \sum_{k=1}^{B} Var\left(P_k\right) = Br.$$

Notice that X is a binomial random variable, with parameters B and p, if there are no multiple countings — that is, if necessarily each $P_k = 0$ or 1. Moreover, if X is sum of independent and identically distributed piece counts, then the central limit theorem implies approximate normality — $X \sim \mathcal{N}(Bp, Br)$ approximately, if the batch size B is large.

Similarly, a second count Y is approximately normal, $\mathcal{N}(Bq, Bs)$, where q and s are the piecewise expectation and variance for the process Y.

Thus, discrepancy $D = X - Y$ is approximately normal, $\mathcal{N}(Bp - Bq, v)$, where variance

$$\begin{aligned} v &= Var(X - Y) \\ &= Var(X) + Var(Y) - 2Cov(X, Y) \\ &= Br + Bs - 2Cov(X, Y) \\ &= B[r + s - 2\rho(X, Y)\sqrt{rs}]. \end{aligned}$$

In particular, the random walk model implies that discrepancy between two counting processes would be *symmetric*, with variance proportional to batch size.

5.3. PERMUTATION PROCEDURES FOR NON-IDENTICALLY DISTRIBUTED OBSERVATIONS

Three approaches already considered for assessing median discrepancy in the independent and identically distributed case are: the Student t test, the sign test, and Wilcoxon signed rank test — and the three corresponding confidence interval procedures. As noted in Chapter 3, symmetry implies that all three approaches are permutation procedures corresponding to distinct score functions. It is well known that the two nonparametric approaches, sign and signed rank procedures, remain valid for non-identically distributed observations. (The sign test and corresponding confidence interval do not

even require symmetry.) [Pratt and Gibbons (1981, p.87 and p.155)]. Efron (1969) also showed that the t test, and hence the t confidence interval procedure, remain valid (and conservative) for non-identically distributed symmetric observations. Thus, under the symmetry assumption, all three approaches still can be applied for non-identically distributed observations.

But relative performance criteria to choose among these permutation procedures have been developed (see Chapter 4) only for the independent and identically distributed case.

5.4. MIXTURE SAMPLING APPROXIMATION TO NON-IDENTICALLY DISTRIBUTED DATA

In order to make the relative performance criteria applicable, one might hope to generalize the criteria for non-identical distributions; unfortunately, there is no clear generalization yet. An alternative, developed below, involves modelling (or approximating) a sequence of non-identically distributed observations by an independent and identically distributed sequence from a mixture.

Suppose that the data $D_1, D_2, \cdots, D_n$ combine m observations from one distribution and $n-m$ observations from a second distribution. That is,

suppose that $D_1, D_2, \cdots, D_n$ are an arbitrary permutation of $D'_1, D'_2, \cdots, D'_m$, $D''_{m+1}, D''_{m+2}, \cdots, D''_n$, where $D'_1, D'_2, \cdots, D'_m$ are m independent and identically distributed observations from F' and $D''_{m+1}, D''_{m+2}, \cdots, D''_n$ are $n-m$ independent and identically distributed observations from F''. Then the combined data are *not* identically distributed; but, with high probability, the data look like an independent and identically distributed sequence drawn from a mixture of F' and F'' with proportions $\frac{m}{n}$ and $\frac{n-m}{n}$, respectively. For sampling from a mixture, $\frac{m}{n}$ would be the *expected* rather than the *actual* fraction from F' (with the distinction vanishing as n gets large and the observed proportion converges to its expected value); and for mixture sampling, the permutation of the $D_1, D_2, \cdots, D_n$ would be rigorously *random* rather than *arbitrary*.

But, if it is difficult in principle to distinguish whether independent data $D_1, D_2, \cdots, D_n$ are the product of simple random sampling from a mixture or of another sampling process involving non-identical distributions (with common point of symmetry), then it seems reasonable to base estimation of the symmetry point on procedures appropriate for the simple mixture sampling. Obviously, this idea generalizes from a mixture of two symmetric distributions to a mixture with three, four, or many components. A sufficiently rich mixture could approximate the random walk model for

discrepancies between counting processes.

Fisher (1955) questioned whether real data ever correspond to "repeated sampling from the same population" and described that assumption as one of the "products of the statistician's imagination". Fisher might have used conditioning (on the observed proportions, etc.) to argue that inference for independent and identically distributed mixture sampling should be the same as for a permutation of non-identically distributed observations. On the other hand, our perspective regards independent and identically distributed sampling from a mixture as a mathematically tractable approximation to the general case.

Notice that the confidence interval based on the mixture would be conservative, because the variance of data from a mixture distribution is greater than that of corresponding data from deterministically non-identical distributions. Suppose U is distributed as a mixture of k symmetric distributions with densities $f_{U_1}, f_{U_2}, \cdots, f_{U_k}$, having a common point of symmetry (assume $E(U_i) = 0$, for $i = 1, 2, \cdots, k$, without loss of generality) and with weights (or expected proportions) $\omega_1, \omega_2, \cdots, \omega_k$, where $0 \leq \omega_i \leq 1$ and $\sum_{i=1}^{k} \omega_i = 1$. Then the probability density function of U, is given by

$f_U(u) = \sum_{i=1}^{k} \omega_i f_{U_i}(u)$. It follows that

$$Var(U) = \int_{-\infty}^{\infty} u^2 f_U(u)du = \int_{-\infty}^{\infty} u^2 \sum_{i=1}^{k} \omega_i f_{U_i}(u)du = \sum_{i=1}^{k} \omega_i \int_{-\infty}^{\infty} u^2 f_{U_i}(u)du$$
$$= \sum_{i=1}^{k} \omega_i \sigma_i^2,$$

where $\sigma_i^2 = Var(U_i) = \int_{-\infty}^{\infty} u^2 f_{U_i}(u)du$. And suppose that V comes from a fixed permutation of independent data from k non-identical densities $f_{U_1}, f_{U_2}, \cdots, f_{U_k}$ with actual proportions $\omega_1, \omega_2, \cdots, \omega_k$. Then $V \sim \sum_{i=1}^{k} \omega_i U_i$, and

$$Var(V) = Var\left(\sum_{i=1}^{k} \omega_i U_i\right) = \sum_{i=1}^{k} \omega_i^2 Var\,(U_i) = \sum_{i=1}^{k} \omega_i^2 \sigma_i^2 \leq \sum_{i=1}^{k} \omega_i \sigma_i^2 = Var(U).$$

5.5. SUMMARY GUIDE TO ESTIMATION OF MEDIAN DISCREPANCY FOR REAL DATA

Monte Carlo simulations have been used to compare performances of the three permutation procedures for particular mixtures (using the performance criteria of Chapter 4). Note that the Wilcoxon signed rank methods perform well for a great variety of normal contaminations or mixtures.

In summary, the following steps are recommended for estimation of median discrepancy for real data.

- Use graphical methods, especially, the average-difference plot for first evaluation of simple assumptions and models.

- If the data clearly are not symmetric, then sign procedures may be the only valid option.

- If, however, we can regard the data as symmetric, then check whether discrepancy variance seems to be constant or a function of the size of measurement.

- If the average-difference plot is consistent with constant variance and independent and identically distributed data, then use normal probability plot, tail-weight statistic, etc., to decide whether the distribution has light, moderate, or heavy tails — and accordingly choose among the permutation procedures (t, sign, or signed rank).

- If discrepancy variance is not constant, but increases with the magnitude measured — especially for discrepancies between integer counts — consider mixture models.

- The signed rank confidence interval is robust in senses noted above. But if length of the signed rank confidence interval greatly exceeds

length of the t or sign confidence interval, then it may be worthwhile to consider (by fresh Monte Carlo simulations) specialized non-normal mixture models.

6. DISCREPANCY OR DISCREPANCY *RATE*?

In contexts where the items measured range from very small to very large magnitudes, it is often preferable to express discrepancy as a *rate*: for instance, a discrepancy of 50 for a shipment of size 100 is very different in importance from the same amount of discrepancy for a shipment of size 5000. More importantly, since in some cases the items being measured are not randomly sampled, but arbitrarily or intentionally chosen over a wide range, a discrepancy *rate* may be a more relevant comparison or more intrinsic characterization to describe the relative bias of two measuring processes. This chapter discusses discrepancy in this relative sense.

As in previous chapters, suppose $X_i = \mu_i + \varepsilon_i$ and $Y_i = \mu_i + \delta_i$, for true magnitude μ_i and random errors ε_i and δ_i. Then the discrepancy is

$$D_i = X_i - Y_i,$$

and we may denote the discrepancy *rate* by

$$R_i = \frac{D_i}{\mu_i}.$$

It follows that

$$E(R_i) = E\left[\frac{D_i}{\mu_i}\right] = \frac{E(D_i)}{\mu_i}, \text{ and}$$
$$Var(R_i) = Var\left[\frac{D_i}{\mu_i}\right] = \frac{Var(D_i)}{\mu_i^2}.$$

In particular, if $Var(D_i)$ is directly proportional to magnitude μ_i, then $Var(R_i)$ is inversely proportional to μ_i.

Thus, if μ_i were known, inference on discrepancy would lead easily to inference on discrepancy rate.

But, of course, μ_i is *not* known — otherwise there would be no need for X_i, Y_i measurements. One natural solution is to use $\hat{\mu}_i = \frac{X_i + Y_i}{2}$ in place of the unknown μ_i. This leads to a practical issue: how relevant is $\hat{R}_i = \frac{D_i}{\hat{\mu}_i}$ to inference about $\frac{D_i}{\mu_i}$?

Suppose $\frac{|X - Y|}{\mu} = |R| \ll 1$, or, equivalently, $|X - Y| \ll \mu$. Then small perturbations in the denominator do not substantially alter the ratio. In practice, measurement errors should be small relative to the magnitude measured; and difference between two small errors should be *very* small. Hence the magnitudes in numerator and denominator of the discrepancy rate are so different that uncertainty in the denominator (due to estimation) usually is irrelevant.

7. APPLICATIONS: EXAMPLES OF DISCREPANCY DATA

Several specific contexts for assessing agreement are presented in this chapter. For some examples, discrepancy data are analyzed in detail; for other examples, we just describe the context and note whether discrepancy variance is proportional to the measured magnitude or to its square, etc.

EXAMPLE 7.1: Counting logs.

This thesis was motivated, in part, by certain questions about counting and "scaling" (measuring volumes) of logs in the British Columbia forest industry. Evidence included data on certain shipments of logs that were counted and "scaled" twice: first at a central facility and again at various destinations. From paired counts, discrepancies can be found by subtraction. Median or relative bias of discrepancy rate was relevant to financial claims.

In particular, two data sets are considered here: 166 batches of logs processed prior to change of the central facility's counting and scaling procedure in October 1981 ("old" logs), and 93 batches of logs after October 1981 ("new" logs). These data are presented in Data Sets 1.1 and 1.2. Generally speaking, these data show that the source counts tended to be slightly below the destination counts in the "old" period, but slightly above

the destination counts in the "new" period. The "old" log discrepancies range roughly from -200 to 120, with more negative than positive; while the "new" log discrepancies range roughly from -150 to 225, with positive and negative discrepancies more or less balanced. See the average-difference plots in Figures 3 and 4.

Several parties were interested in discrepancy rates for these data; and because of the substantial dollar amounts involved, point estimation and confidence intervals were of great concern (while hypothesis testing was less important). Professor Glick's analyses considered data subsets determined by calendar year, species of log, and so on, as well as the "old" and "new" data overall.

The scatter plots of source counts versus destination counts for the "old" logs and "new" logs are given in Figures 1 and 2, respectively. Both scatter plots indicate high correlation — with data points tightly along a straight line. But recall that high correlation does not necessarily suggest strong agreement in the present context; see Chapter 2.

Figure 1 seems to suggest heteroscedasticity — that the amount of variability of discrepancy increases with the batch sizes for the "old" logs,

but the trend in not clear. This phenomenon is even less obvious in Figure 2 for the "new" logs. However, heteroscedasticity is clear in two average-difference plots, Figures 3 and 4, respectively — the "shotgun" pattern, spreading out like a root function, suggests that the variance of discrepancy is proportional to the batch size. This proportionality phenomenon can be displayed more clearly in a plot of difference $(X - Y)$ versus square root of the average $\left(\sqrt{\frac{X+Y}{2}}\right)$; see Figure 5 and 6. The context (counting) suggests a "random walk" model that is compatible with these graphical results and that could be approximated by a mixture of normals which differ only in their variances; see Chapter 5.

Since the batch sizes cover a large range (roughly from 100 to 3000) and since the batches are not randomly, but arbitrarily chosen, it would be preferable to consider discrepancy rate rather than the simple difference; see Chapter 6. The normal probability plot of the "old" discrepancy rates, provided in Figure 7, exhibits a certain degree of linearity (except for one outlier, with discrepancy rate roughly 28%) although, as just noted, the average-difference plots, Figures 5 and 6, indicate that the data are not identically normally distributed. This normal probability plot is fairly similar to plots for data simulated from the scale mixtures of four and five normals (mixtures discussed in Chapter 4, although probability plots for simulated

mixture sampling have not been shown).

Hence, for these data (and for subsets of these data), the Wilcoxon signed rank procedure is preferable to its competitors, based on the simulation results in Chapter 4. This preference is more-or-less compatible with the tail-weight statistic used by the adaptive procedure mentioned in Appendix III: the statistic $Q^* = 3.06$, while 2.92 is the suggested boundary value between "moderate" and "heavy" tails. For all 166 "old" batches of wood, the Wilcoxon signed rank interval, with 95% confidence, estimates that the median discrepancy rate (source count minus destination count) is negative, with magnitude interval 1.28% to 2.83%; see Table 7.

Student t and sign confidence intervals also were calculated for the 166 "old" batches of logs. Tables 7 and 8 show that length ratios of these intervals relative to Wilcoxon signed rank interval are very similar to corresponding interval length ratios for scale mixture of four or five normals studied in Chapter 4. These ratio results provide further support for applying to these data the "random walk" model and the corresponding scale mixture approximation.

The normal probability plot of the "new" discrepancy rates, provided in

Figure 8, shows that the five smallest and the two largest rates are potential outliers, which would be deleted for further analysis. The probability plot of the remaining rates, given in Figure 9, indicates a complicated mixture, with some skewness to the right. Although simulation results show that the Wilcoxon signed rank procedures are robust over a wide variety of distributions, asymmetry and heavy-tailedness (classified by the Q^* tail-weight statistic) make the sign procedure a good choice for the "new" logs.

For 86 batches of "new" logs (after deleting outliers), the sign confidence interval, with 96% confidence, estimates that the median discrepancy rate (source count minus destination count) is 0% to 0.34%; see Tables 9 and 10. Note that for interval estimation using the exact sign procedure with sample size $n = 86$, an exact confidence level 95% is not attainable, and 96% is as close as possible.

EXAMPLE 7.2: Fuse-burning times.

Grubbs (1948) gave burning times (in seconds) of 30 powder train fuses reported by three observers, say A, B, and C. Since one burning time for observer B was lost, this example only considers data for observers A and C, whose times are provided in Data Set 2. Scatter plot and average-difference

plot are provided in Figures 10 and 11, respectively.

Notice that although the correlation between burning times recorded by observers A and C is high (0.99), the average-difference plot does show some systematic disagreement between the two observers.

Grubbs (1948) used a components-of-variance model, assuming that errors are unrelated to the times measured and are identically distributed for all observers. He partitioned variation into two components: due to fuse variation, and due to observer error. However, the average-difference plot, showing a "shotgun" pattern, suggests that the standard deviation of discrepancy may be proportional to the size of measurement, and hence that the validity of Grubbs' assumption is questionable. Indeed, Grubbs did notice that "errors of measurement (e) in some cases increase with increasing magnitude of the characteristic measured (x)". But he assumed that "x and e are sufficiently independent to insure that limited variations in x are not reflected in the errors of measurement".

Such assumption has often been made in literature of agreement problems. This example draws our attention to the need for considering this assumption more seriously and for finding appropriate methods when the

assumption does not hold.

EXAMPLE 7.3: Systolic blood pressure readings.

Systolic blood pressures (in mm Hg) measured by two different methods on 25 patients were used in a textbook example of correlation by Daniel (1983). This example also was discussed by Altman and Bland (1983); the data are listed in Data Set 3. Scatter plot and average-difference plot are given in Figure 12 and 13, respectively.

Note that although the correlation coefficient between readings by the two methods is high (approximately 0.95), this does not imply agreement between the methods, in the sense of low relative bias; in fact, disagreement is clear in the average-difference plot. The average-difference plot also exhibits a "shotgun" pattern and hence calls into question the assumption of error distribution with constant variance.

EXAMPLE 7.4: Spinal curvature — angular data.

Spinal curvature, which is often used as a clinical assessment of scoliosis, can be described by two angles, viz., the Ferguson angle and the Cobb angle. The data in Data Set 4 come from a study comparing these two angles

for $n = 26$ patients [Robinson and Wade (1983)]. Predictability of one angle from the other angle seems to be the primary interest, but relative bias also would be interesting.

Scatter plot and average-difference plot are given in Figures 14 and 15, respectively. This average-difference plot exhibits a pattern like an ordinary X, Y scatter plot and differs from all the average-difference plots considered above. This implies that error variances for Ferguson and Cobb angle measurements are not equal; see Chapter 2 and Altman and Bland (1983). Further study of *replicated* Ferguson measurements and *replicated* Cobb measurements, on the same patients, likely would show greater reliability (higher correlation) for one method relative to the other, and hence may suggest practical preference for one method.

Disagreement between the two angles is clear (even though correlation coefficient is 0.95) — Cobb angle is uniformly greater than the corresponding Ferguson angle. Moreover, relative bias of discrepancy obviously increases with the size of the angle, as shown in the average-difference plot. Hence it is not clear that any estimation methods considered above would be appropriate. And when the Cobb angle is regressed on the Ferguson angle, as by Robinson and Wade (1983), interpretation of the intercept is

questionable. Also, since both the Cobb and Ferguson angles are measured with errors, it is inappropriate to apply the usual regression; see Chapter 2. The plot of residuals versus the Ferguson angles, provided in Figure 16, indicates that the residuals do depend on the Ferguson angle.

EXAMPLE 7.5: Oxygen levels for newborn infants.

This data set comes from a study of newborn infants, comparing a "containing" position in a hammock with the supine position, when measuring respiration [Bottos, et al. (1985)].

Oxygen levels (pressure in mm Hg) of 50 babies measured in both positions are given in Data Set 5. Scatter plot and average-difference plot are given in Figures 17 and 18, respectively. The average-difference plot shows that relative bias between oxygen measurements in the two positions may be small (differences approximately symmetric around zero horizontal) but variation of discrepancy obviously increases with the oxygen level. (Note the substantial difference in oxygen for baby number 14 — 55.42 mmHg in supine position, 108.92 mmHg in hammock position.) These observations suggest usage of a nonparametric or robust confidence interval procedure.

EXAMPLE 7.6: Tobacco moisture content.

These data come from a study of two electrical devices, say A and B, which measure the moisture content of tobacco. Data of 15 tobacco samples are listed in Data Set 6 (adapted from a B.Sc. Special Examination, University of London; no unit specified for "moisture content"). Scatter plot and average-difference plot are shown in Figure 19 and 20, respectively.

Again, although the correlation is high (0.996), the average-difference plot suggests that variance of discrepancy increases with the moisture content; however, the trend is not very clear, possibly because of small sample size.

Table 1: **Efficacies and Pitman asymptotic relative efficiency (ARE) comparisons of Student t (T), sign (S), and Wilcoxon signed rank (W) procedures.** Numeric efficacies are for familiar densities standardized so that efficacy is 1 for the sign procedure. Hence, the entries also are Pitman asymptotic efficiencies relative to the sign procedure. For example, $ARE(T,S) = 1.57$ for normal distribution. [Adapted from Pratt and Gibbons (1981, p.384)]; see also Appendix II.

Distributions	W	T	S
Normal (0,$2/\pi$)	1.50	1.57	1.00
Uniform (-1,1)	3.00	3.00	1.00
Cauchy (0,$2/\pi$)	0.75	0.00	1.00

Table 2: **Pitman asymptotic relative efficiency (ARE) comparisons of Student t (T), sign (S), and Wilcoxon signed rank (W) procedures (results using theoretical efficacies).** Relative efficiencies of procedures for normal, Cauchy, and uniform distributions do not depend on location and scale parameters of the distributions; see also Table 1. The listed distributions include standard normal and Cauchy; uniform distribution over $(-1, 1)$; and normals mixed or contaminated: N.Mix(1 : 4) denotes an equal-proportions mixture of normals $\mathcal{N}(0, i^2)$ for $i = 1, 2, 3, 4$; CN(2;5) denotes 5% of $\mathcal{N}(0, 2^2)$ contaminating standard normal $\mathcal{N}(0, 1)$; etc.

	Pitman ARE		
Distributions	$W : S$	$T : S$	$W : T$
Normal	1.5000	1.5708	0.9549
Uniform	3.0000	3.0000	1.0000
Cauchy	0.7500	0.0000	∞
N.Mix(1:5)	0.9568	0.6847	1.3973
N.Mix(1:4)	1.0317	0.7721	1.3363
CN(2;5)	1.4658	1.4369	1.0202
CN(2;1)	1.4930	1.5404	0.9692
CN(4;5)	1.4177	0.9689	1.4632
CN(4;1)	1.4832	1.3866	1.0696
CN(10;5)	1.3803	0.2895	4.7687
CN(10;1)	1.4755	0.8037	1.8358

Table 3: **Asymptotic ($n \to \infty$) ratios of lengths and squared lengths of confidence intervals ($\sqrt{1/\mathrm{ARE}}$ and 1/ARE, respectively, where ARE's are Pitman asymptotic relative efficiencies) for Student t (T), sign (S), and Wilcoxon signed rank (W) procedures (results using theoretical efficacies).** Asymptotic ratios for normal, Cauchy, and uniform distributions do not depend on location and scale parameters of the distributions. For details of other listed distributions, see Table 2.

	Lengths Ratio $= \sqrt{1/ARE}$			*Sq. Lengths Ratio* $= 1/ARE$		
Distributions	$W:S$	$T:S$	$W:T$	$W:S$	$T:S$	$W:T$
Normal	0.8165	0.7979	1.0233	0.6667	0.6366	1.0472
Uniform	0.5773	0.5773	1.0000	0.3333	0.3333	1.0000
Cauchy	1.1547	∞	0.0000	1.3333	∞	0.0000
N.Mix(1:5)	1.0224	1.2085	0.8460	1.0452	1.4604	0.7157
N.Mix(1:4)	0.9845	1.1381	0.8650	0.9693	1.2952	0.7483
CN(2;5)	0.8260	0.8343	0.9901	0.6822	0.6960	0.9802
CN(2;1)	0.8184	0.8057	1.0158	0.6698	0.6492	1.0318
CN(4;5)	0.8399	1.0159	0.8267	0.7054	1.0321	0.6835
CN(4;1)	0.8211	0.8492	0.9669	0.6742	0.7212	0.9349
CN(10;5)	0.8512	1.8587	0.4579	0.7245	3.4546	0.2097
CN(10;1)	0.8232	1.1154	0.7380	0.6777	1.2442	0.5447

Table 4: **Comparisons of T, S, and W confidence intervals: average ratios of interval lengths and squared lengths (Monte Carlo simulation results).** Each table entry is based on 1000 samples with $n = 32$, and all intervals have nominal 95% confidence; see Table 2 for descriptions of the listed distributions. For example, among 1000 standard normal samples, 0.8900 was the average $W : S$ length ratio (dividing length of the Wilcoxon interval by length of the interval corresponding to the sign test, for each sample).

	Mean of Lengths Ratio			*Mean of Sq. Lengths Ratio*		
Distributions	$W : S$	$T : S$	$W : T$	$W : S$	$T : S$	$W : T$
Normal	0.8900	0.8704	1.0314	0.8345	0.8103	1.0684
Uniform	0.7233	0.6614	1.0905	0.5598	0.4648	1.1916
Cauchy	1.3522	11.7286	0.3647	2.0823	1478.381	0.1941
N.Mix(1:5)	1.0636	1.1843	0.9164	1.1968	1.5240	0.8510
N.Mix(1:4)	1.0402	1.1485	0.9258	1.1485	1.4510	0.8691
CN(2;5)	0.8921	0.8930	1.0105	0.8374	0.8530	1.0276
CN(2;1)	0.8866	0.8715	1.0263	0.8280	0.8103	1.0580
CN(4;5)	0.9123	1.0487	0.9116	0.8746	1.2300	0.8574
CN(4;1)	0.8909	0.9094	0.9992	0.8360	0.8953	1.0105
CN(10;5)	0.9326	1.7199	0.6795	0.9170	3.9835	0.5413
CN(10;1)	0.8941	1.0847	0.9243	0.8421	1.5144	0.9026

Table 5: **Comparisons of T, S, and W confidence intervals: ratios of standard deviations of interval lengths (Monte Carlo simulation results).** Each entry is based on 1000 samples with $n = 32$; see Table 2 for descriptions of the listed distributions. For example, the first entry shows that SD of Wilcoxon interval lengths, divided by SD of interval lengths for the sign procedure, gives a ratio 0.4696, for normal samples.

	Ratio of SD of Lengths		
Distributions	$W(SD) : S(SD)$	$T(SD) : S(SD)$	$W(SD) : T(SD)$
Normal	0.4696	0.4156	1.1299
Uniform	0.2428	0.2040	1.1904
Cauchy	1.2385	110.1337	0.0112
N.Mix(1:5)	0.6479	0.6418	1.0094
N.Mix(1:4)	0.6451	0.6408	1.0067
CN(2;5)	0.4787	0.4960	0.9651
CN(2;1)	0.4538	0.4188	1.0835
CN(4;5)	0.5361	1.1005	0.4871
CN(4;1)	0.4608	0.6441	0.7155
CN(10;5)	0.6028	3.5477	0.1699
CN(10;1)	0.4696	2.0320	0.2311

Table 6: **Confidence interval coverages and length comparisons (Monte Carlo simulation results).** For each distribution listed below (all are symmetric about zero; see Table 2 for descriptions) 1000 samples were simulated, each with sample size $n = 32$; and, for each sample, confidence interval estimates of distribution median were calculated using Student t, sign, and Wilcoxon signed rank procedures (denoted respectively as T, S, W) with nominal confidence level 95%. Table entries: a) for each of the three procedures, the percentage of samples (from each distribution) for which the calculated intervals included zero; and b) the percentage of samples for which the length of one interval was shorter than another ($T < S$, etc.).

	% Coverage			*% Shorter Length*		
Distributions	T	S	W	T<S	T<W	W<S
Normal	94.6	94.4	95.3	77.2	73.1	75.9
Uniform	95.5	93.8	96.1	95.5	96.8	92.5
Cauchy	98.0	93.8	96.1	1.0	0.6	17.7
N.Mix(1:5)	94.3	95.0	94.7	33.7	23.2	47.0
N.Mix(1:4)	95.5	94.9	94.6	39.5	27.1	51.8
CN(2;5)	94.6	94.4	95.0	74.1	63.3	77.2
CN(2;1)	94.7	94.4	95.4	77.9	70.6	78.4
CN(4;5)	95.6	94.4	94.9	54.2	39.3	74.2
CN(4;1)	95.3	94.4	95.2	72.4	62.7	78.1
CN(10;5)	96.9	94.4	94.6	28.1	21.6	71.1
CN(10;1)	96.1	94.4	95.3	62.1	55.0	78.1

Table 7: **(Example 7.1). Point estimate and confidence interval for median discrepancy rate of "old" logging counts ($n = 166$ batches) using T, S, and W procedures. See Data Set 1.1.**

Procedure	Pt Est	Confidence Level	Interval Endpoints	Length
T	−0.0269	95.00%	(−0.0356,−0.0182)	0.0174
S	−0.0225	94.80%	(−0.0283,−0.0128)	0.0154
W	−0.0255	95.00%	(−0.0338,−0.0177)	0.0161

Table 8: **(Example 7.1). Ratios of lengths and squared lengths for confidence intervals in Table 7.**

Procedures	Length Ratio	Squared Length Ratio
W:S	1.0446	1.0911
T:S	1.1291	1.2748
W:T	0.9251	0.8559

Table 9: **(Example 7.1). Point estimate and confidence interval for median discrepancy rate of "new" logging counts ($n = 86$ batches — with 7 outliers deleted) using T, S, and W procedures. See Data Set 1.2.**

Procedure	Pt Est	Confidence Level	Interval Endpoints	Length
T	0.0038	95.00%	(−0.00244,0.01003)	0.01247
S	0.0005	96.01%	(0.00000,0.00343)	0.00343
W	0.0018	95.00%	(−0.00266,0.00746)	0.01012

Table 10: **(Example 7.1). Ratios of lengths and squared lengths for confidence intervals in Table 9.**

Procedures	Length Ratio	Squared Length Ratio
W:S	2.9459	8.6783
T:S	3.6311	13.1850
W:T	0.8113	0.6582

Data Set 1.1: **(Example 7.1). Source counts and destination counts for 166 batches of "old" logs.**

Batch Number	Source Count	Destination Count
1	1068	1116
2	623	624
3	1655	1644
4	672	683
5	21	19
6	2402	2398
7	551	547
8	1026	1065
9	148	164
10	2489	2503
11	2272	2455
12	850	844
13	409	398
14	729	719
15	1027	1006
16	548	546
17	240	242
18	395	386
19	456	458
20	112	118
21	2437	2614
22	1963	1977
23	587	587
24	2516	2719
25	765	768
26	506	511
27	2161	2209
28	1547	1567
29	867	866
30	638	662

Batch Number	Source Count	Destination Count
31	302	312
32	630	650
33	595	659
34	580	647
35	375	369
36	575	577
37	1172	1200
38	589	608
39	696	690
40	529	539
41	117	113
42	418	468
43	745	762
44	1066	1169
45	1199	1302
46	687	656
47	1233	1240
48	1029	1052
49	846	852
50	519	503
51	1898	1997
52	1475	1438
53	883	957
54	1319	1329
55	887	946
56	1028	1075
57	1176	1208
58	929	925
59	1392	1391
60	544	574

· · · continued below

Data Set 1.1: **(Example 7.1).** ··· continued

Batch Number	Source Count	Destination Count
61	467	503
62	171	186
63	571	584
64	129	123
65	309	311
66	974	1098
67	1105	1267
68	891	871
69	1355	1387
70	1314	1371
71	848	924
72	631	654
73	904	894
74	705	713
75	717	702
76	883	882
77	1349	1347
78	1132	1052
79	870	933
80	894	894
81	1013	1051
82	696	803
83	850	874
84	2436	2501
85	291	289
86	1071	1054
87	865	832
88	1217	1234
89	1074	1112
90	855	872

Batch Number	Source Count	Destination Count
91	1014	1060
92	446	468
93	642	677
94	498	592
95	646	538
96	526	565
97	35	33
98	516	550
99	587	625
100	19	19
101	865	869
102	1065	1092
103	910	934
104	298	322
105	947	947
106	1046	1067
107	405	394
108	644	651
109	668	641
110	610	647
111	1005	1000
112	277	291
113	675	689
114	649	667
115	1433	1596
116	868	908
117	1269	1181
118	900	893
119	277	294
120	364	377

··· continued below

Data Set 1.1: **(Example 7.1).** $\cdots$ continued

Batch Number	Source Count	Destination Count
121	891	915
122	775	777
123	666	630
124	1085	1184
125	1329	1436
126	320	328
127	610	606
128	327	434
129	930	886
130	2445	2616
131	725	638
132	797	945
133	2162	2456
134	835	920
135	1362	1377
136	772	797
137	838	838
138	581	566
139	885	924
140	1411	1301
141	534	538
142	1033	1134
143	1158	1250

Batch Number	Source Count	Destination Count
144	1004	1055
145	1054	1006
146	864	899
147	449	437
148	438	441
149	273	243
150	585	668
151	734	746
152	597	630
153	443	479
154	423	443
155	562	603
156	900	913
157	977	1005
158	1045	1073
159	696	735
160	513	519
161	817	906
162	504	564
163	1201	1280
164	564	573
165	1209	1302
166	236	256

Data Set 1.2: **(Example 7.1). Source counts and destination counts for 93 batches of "new" logs.**

Batch Number	Source Count	Destination Count
1	2833	2622
2	507	508
3	376	382
4	158	160
5	486	485
6	623	638
7	779	768
8	639	661
9	951	950
10	465	482
11	1035	1032
12	841	806
13	641	718
14	670	668
15	534	529
16	590	583
17	720	719
18	1011	1008
19	880	906
20	711	733
21	476	495
22	617	631
23	522	540
24	395	401
25	111	128

Batch Number	Source Count	Destination Count
26	596	594
27	783	762
28	555	587
29	2489	2481
30	2693	2663
31	654	653
32	875	872
33	2344	2392
34	380	384
35	271	267
36	354	352
37	509	582
38	721	603
39	477	480
40	849	838
41	347	363
42	690	668
43	341	322
44	686	699
45	753	763
46	859	859
47	876	902
48	611	603
49	770	777
50	408	408

... continued below

Data Set 1.2: **(Example 7.1).** $\cdots$ continued

Batch Number	Source Count	Destination Count
51	612	599
52	171	168
53	964	1118
54	327	318
55	609	641
56	229	252
57	541	544
58	607	639
59	522	512
60	266	281
61	164	164
62	653	633
63	252	252
64	361	343
65	385	381
66	238	237
67	1504	1392
68	221	208
69	691	648
70	1464	1246
71	862	868
72	613	595

Batch Number	Source Count	Destination Count
73	984	990
74	546	543
75	265	265
76	351	351
77	254	253
78	396	365
79	486	487
80	758	744
81	496	479
82	296	296
83	467	467
84	238	229
85	654	661
86	303	304
87	306	307
88	297	297
89	1881	1791
90	2158	2158
91	295	297
92	2524	2564
93	300	287

Data Set 2: **(Example 7.2). Fuse burning times (seconds) measured by two observers for 30 powder train fuses.** [Grubbs (1948)].

Sample Number	*Observer*	
	A	B
1	10.10	10.07
2	9.98	9.90
3	9.89	9.86
4	9.79	9.70
5	9.67	9.65
6	9.89	9.83
7	9.82	9.79
8	9.59	9.59
9	9.76	9.72
10	9.93	9.92
11	9.62	9.64
12	10.24	10.24
13	9.84	9.86
14	9.62	9.63
15	9.60	9.65

Sample Number	*Observer*	
	A	B
16	9.74	9.74
17	10.32	10.34
18	9.86	9.86
19	10.01	10.03
20	9.65	9.65
21	9.50	9.50
22	9.56	9.55
23	9.54	9.54
24	9.89	9.88
25	9.53	9.51
26	9.52	9.53
27	9.44	9.45
28	9.67	9.67
29	9.77	9.78
30	9.86	9.86

Data Set 3: **(Example 7.3). Systolic blood pressures (mm Hg) by two methods in 25 patients.** [Daniel (1983)].

Patient Number	*Method*	
	I	II
1	132	130
2	138	134
3	144	132
4	146	140
5	148	150
6	152	144
7	158	150
8	130	122
9	162	160
10	168	150
11	172	160
12	174	178
13	180	168
14	180	174
15	188	186
16	194	172
17	194	182
18	200	178
19	200	196
20	204	188
21	210	180
22	210	196
23	216	210
24	220	190
25	220	202

Data Set 4: **(Example 7.4). Spinal curvature (angle, in degrees) by Ferguson method and by Cobb method in 26 patients.** [Robinson and Wade (1983)].

Patient Number	*Method*	
	Ferguson	Cobb
1	73	97
2	66	90
3	60	88
4	50	67
5	48	70
6	47	63
7	45	55
8	43	50
9	43	48
10	40	65
11	40	64
12	38	47
13	37	52
14	37	49
15	36	60
16	36	48
17	33	41
18	30	45
19	30	40
20	29	45
21	29	39
22	28	42
23	28	37
24	27	39
25	27	35
26	21	28

Data Set 5: **(Example 7.5). Cutaneous oxygen levels (mmHg) in 50 newborn infants measured in two positions.** [Bottos, et al. (1985)].

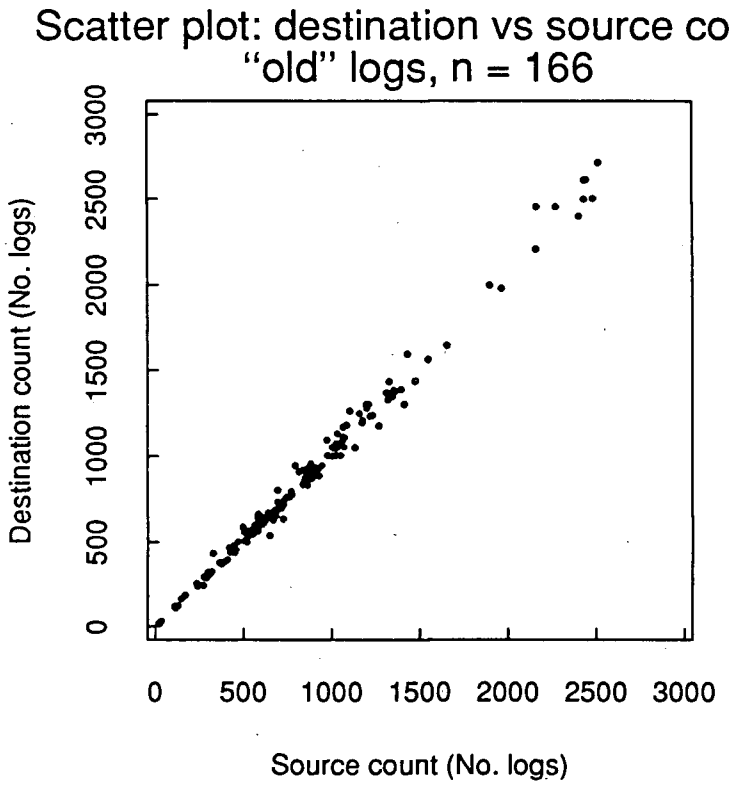
Infant	*Position*	
Number	Hammock	Supine
1	80.67	93.83
2	56.13	69.08
3	95.17	103.58
4	66.42	68.88
5	77.42	67.83
6	55.92	59.50
7	61.79	60.50
8	65.92	68.71
9	65.71	65.54
10	67.30	75.33
11	77.17	69.67
12	71.67	67.13
13	85.00	77.79
14	108.92	55.42
15	52.71	57.59
16	66.00	62.67
17	75.83	78.83
18	66.83	64.04
19	76.04	70.50
20	67.71	56.63
21	72.00	77.21
22	69.96	71.75
23	87.71	72.75
24	82.33	76.38
25	84.63	79.83

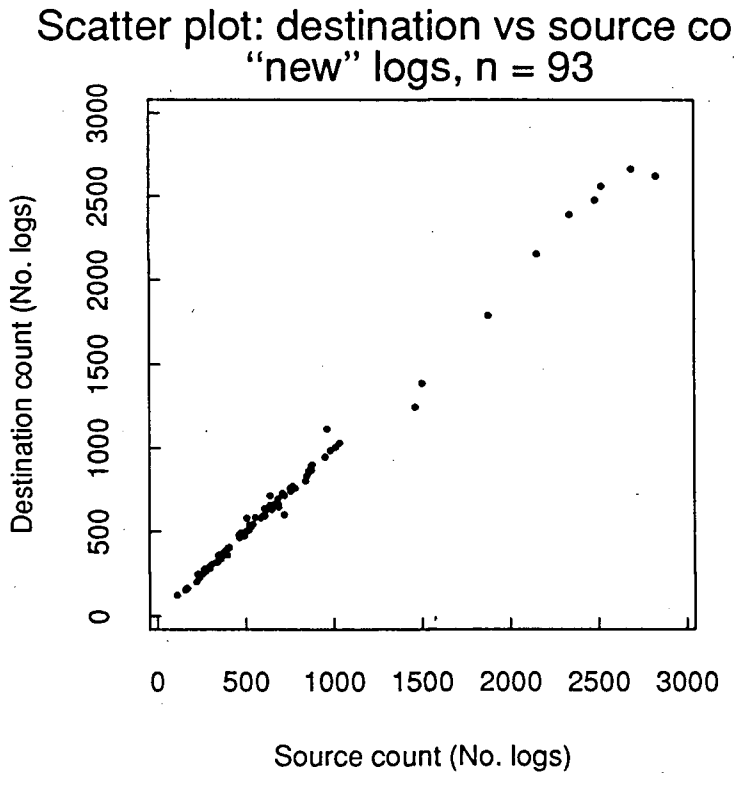
Infant	*Position*	
Number	Hammock	Supine
26	77.46	96.75
27	60.96	54.04
28	74.33	69.46
29	52.67	71.83
30	52.96	58.67
31	71.50	62.88
32	56.96	55.21
33	66.67	59.79
34	67.58	72.75
35	69.92	77.71
36	86.29	85.00
37	55.67	54.33
38	64.25	76.58
39	71.71	75.50
40	71.13	85.83
41	72.63	85.54
42	50.58	87.54
43	49.29	56.88
44	82.83	79.75
45	88.58	80.13
46	58.95	61.96
47	54.17	63.83
48	49.96	50.00
49	80.25	61.17
50	60.96	56.88

Data Set 6: **(Example 7.6). Tobacco moisture content in 15 samples measured by two devices.** [Adapted from a B.Sc. Special Examination, University of London].

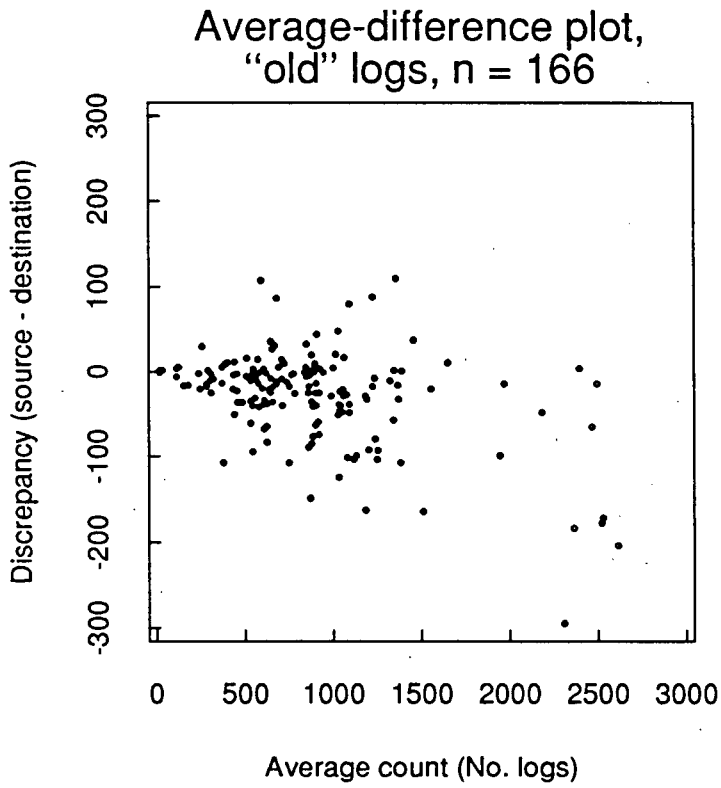
Sample Number	*Device*	
	A	B
1	12.0	10.1
2	12.1	13.5
3	7.5	8.5
4	8.0	9.6
5	16.0	16.8
6	24.5	23.6
7	5.0	4.9
8	47.9	47.8
9	43.1	46.7
10	38.2	38.3
11	69.0	64.8
12	11.8	12.0
13	20.0	17.5
14	57.6	55.2
15	15.0	14.8

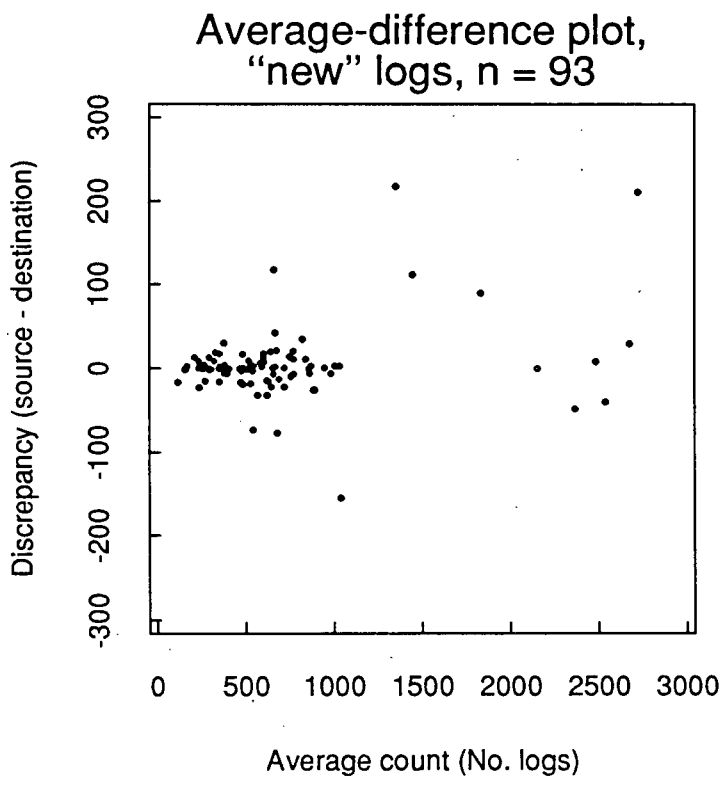
Figures 1 & 2: Counting logs (Data Set 1)



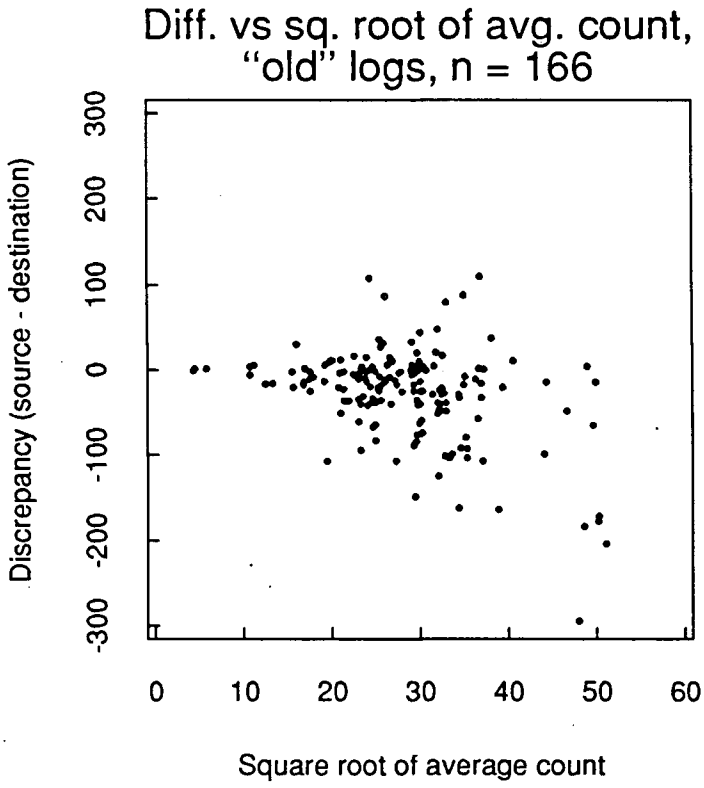


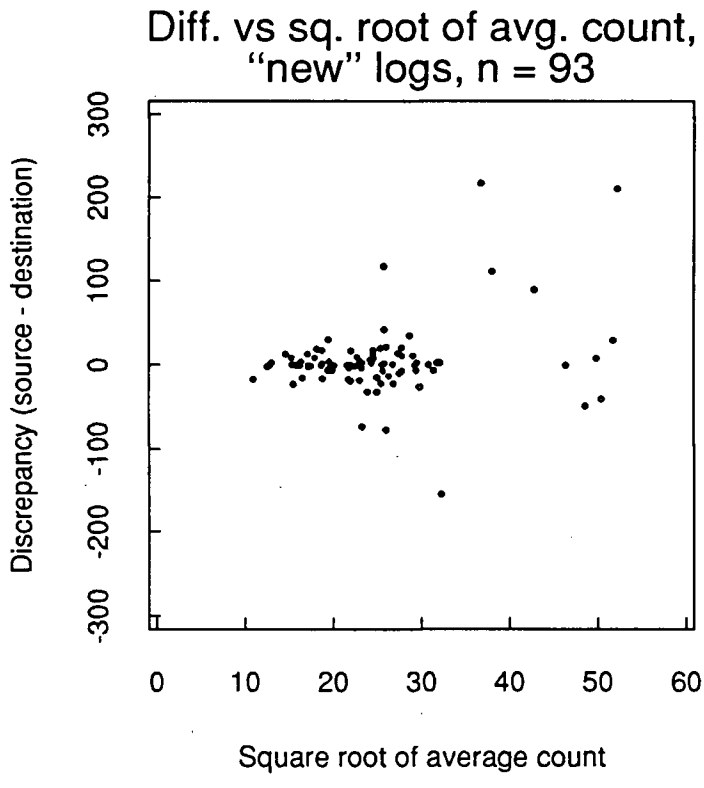
Figures 3 & 4: Counting logs (Data Set 1)



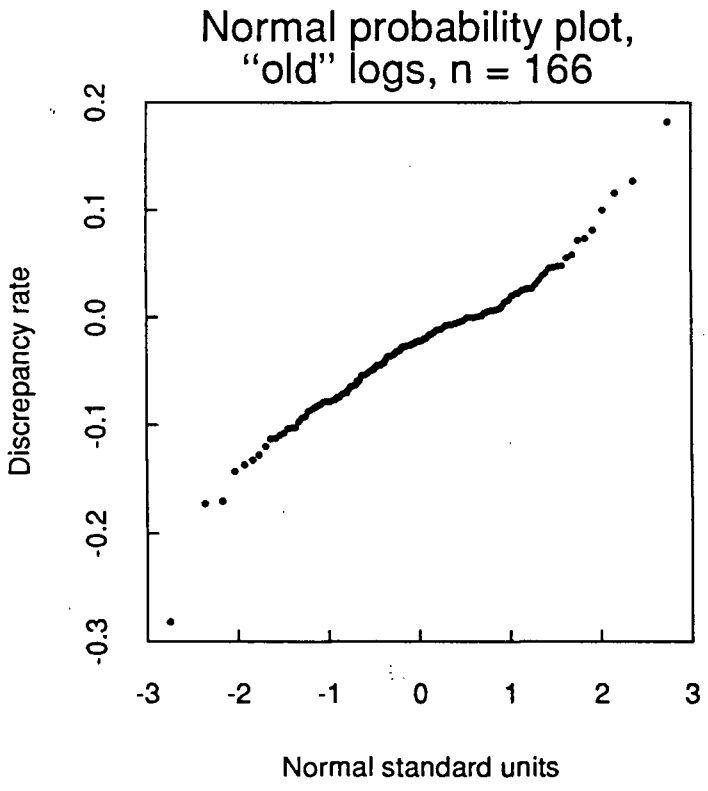


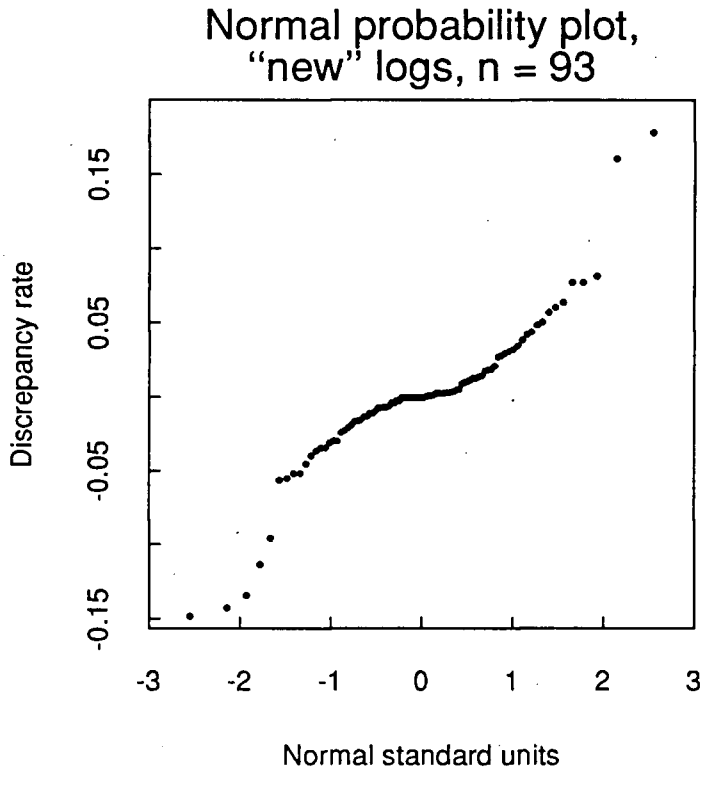
Figures 5 & 6: Counting logs (Data Set 1)



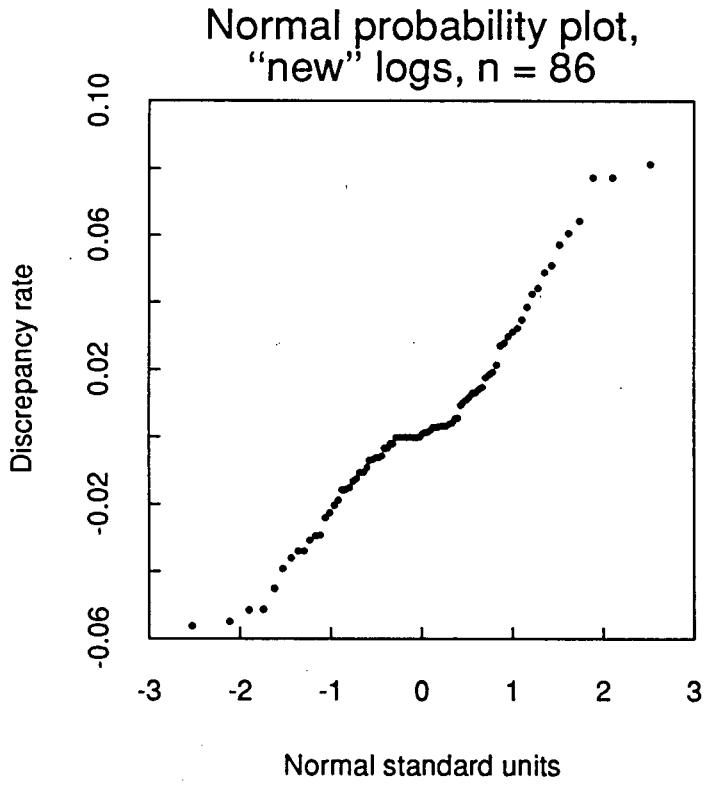


Figures 7 & 8: Counting logs (Data Set 1)



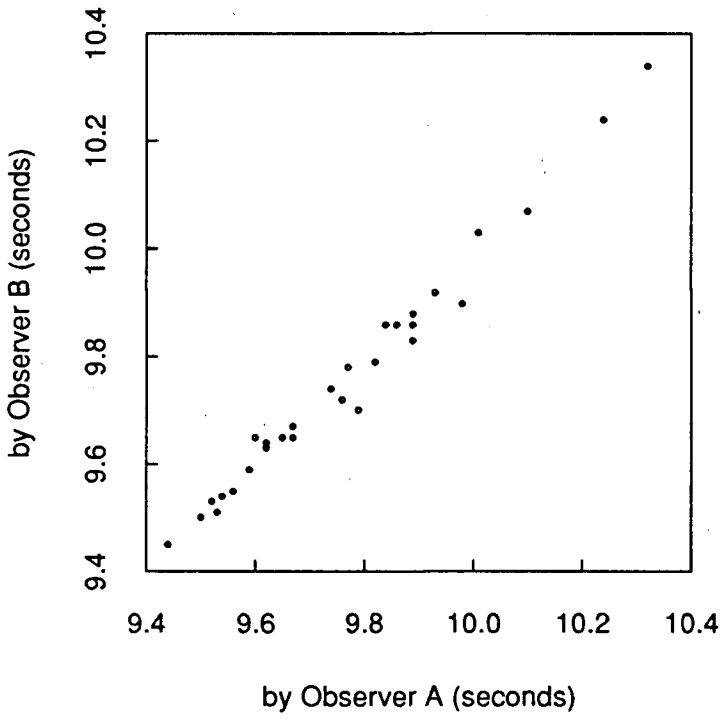


Figures 9: Counting logs
(subset of Data Set 1.2 --- 7 outliers are deleted)

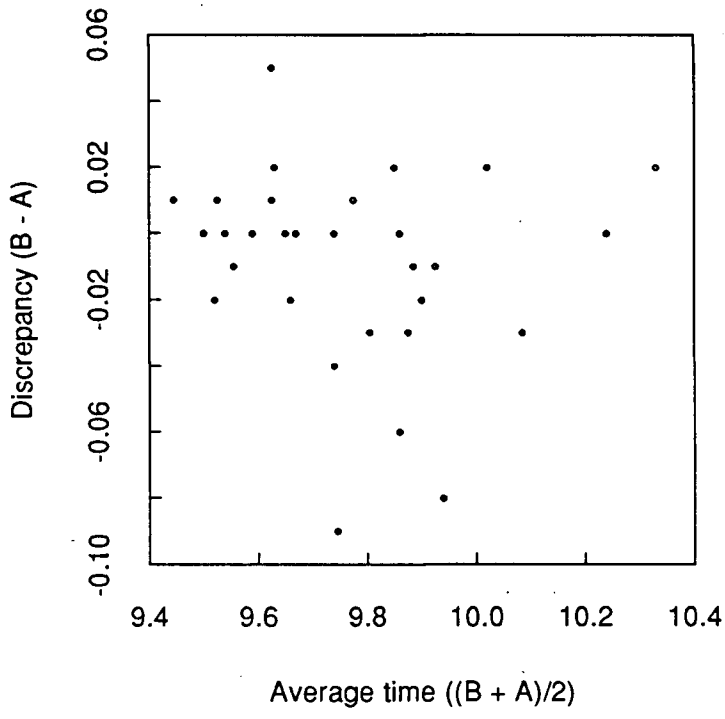


Figures 10 & 11: Fuse burning times (Data Set 2, n = 30)

Scatter plot

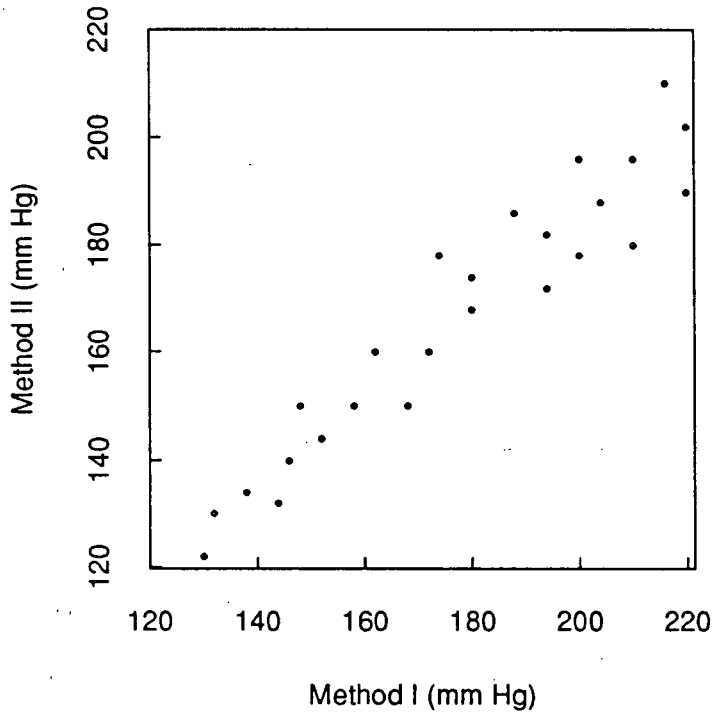


Average-difference plot

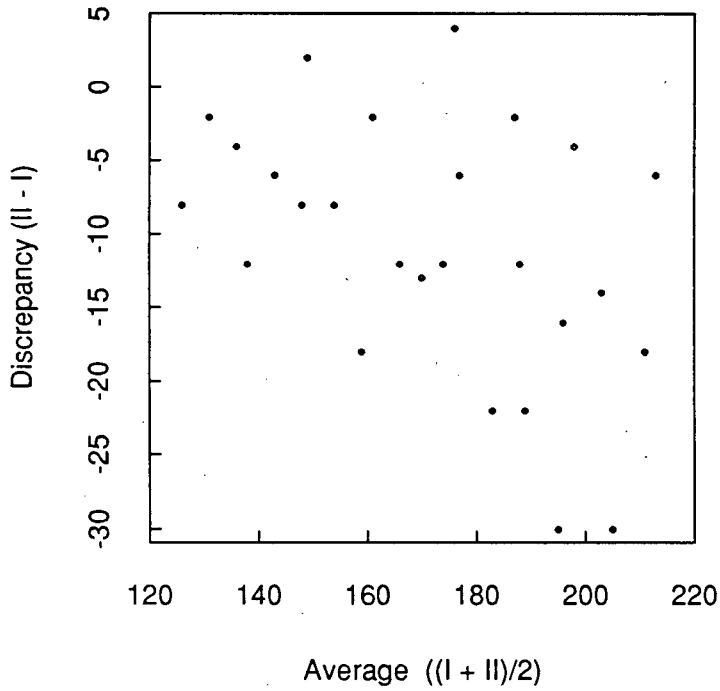


Figures 12 & 13: Systolic blood pressures (Data Set 3, n = 25)

Scatter plot

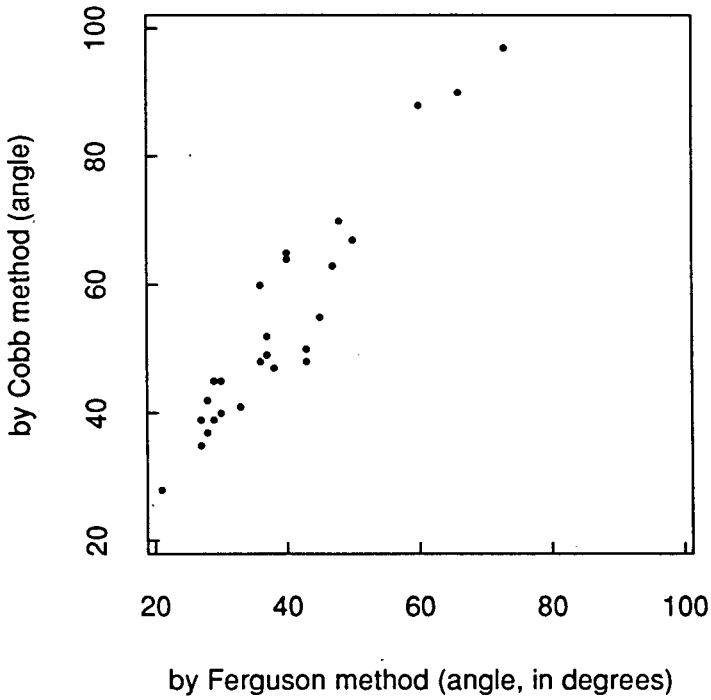


Average-difference plot

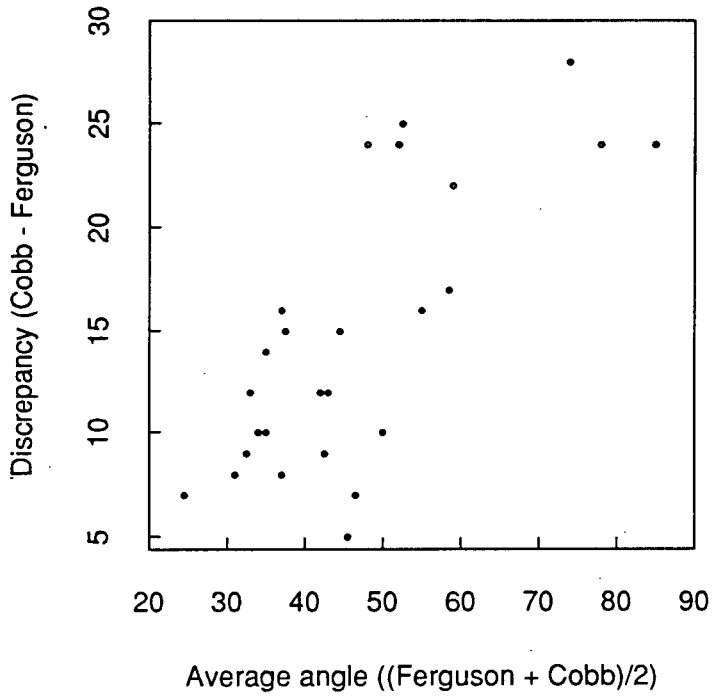


Figures 14 & 15: Spinal curvature (Data Set 4, n = 26)

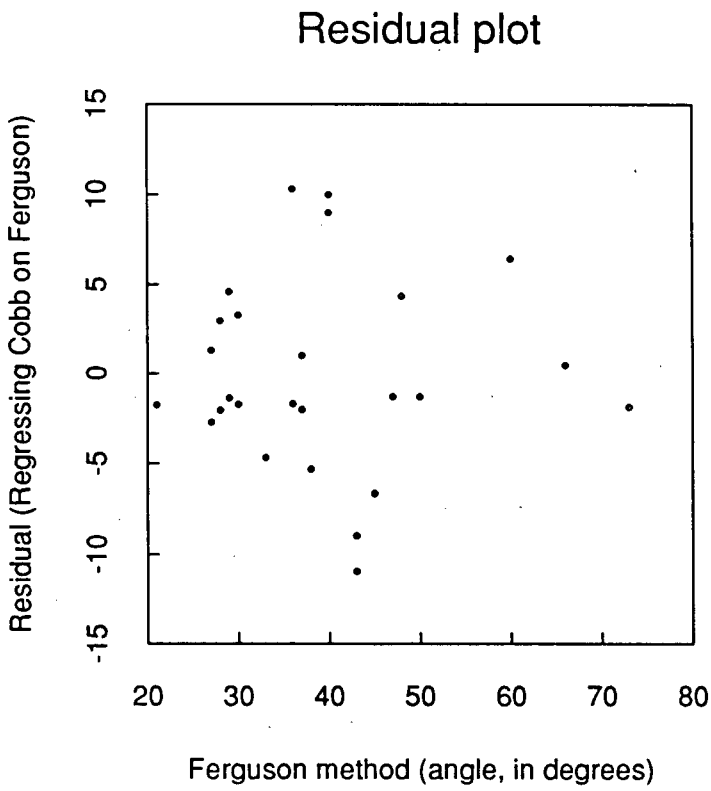
Scatter plot



Average-difference plot

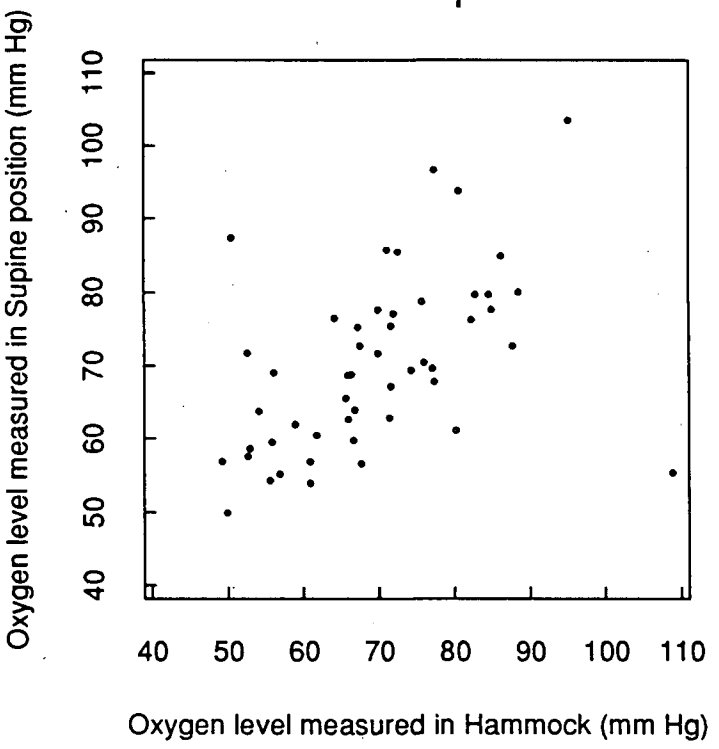


Figures 16: Spinal curvature
(Data Set 4, n = 26)

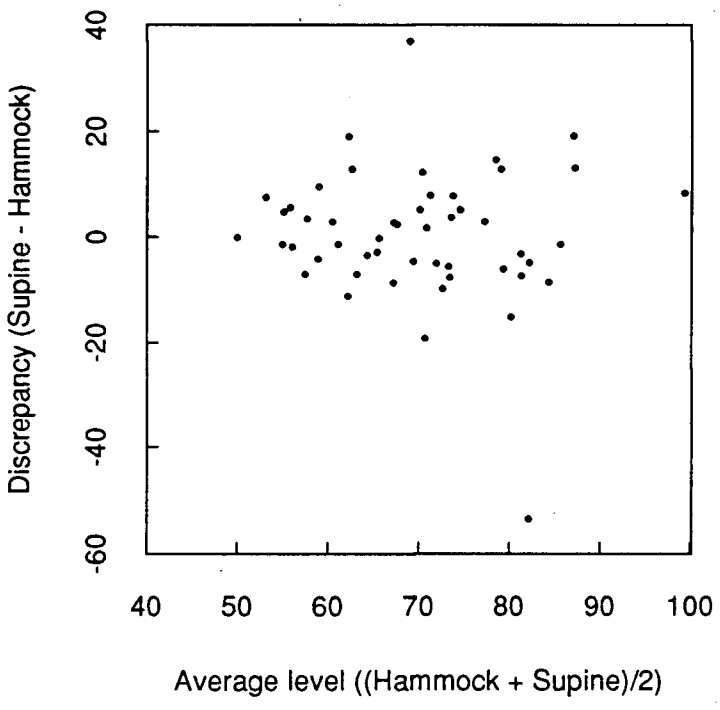


Figures 17 & 18: Oxygen level and position (Data Set 5, n = 50)

Scatter plot

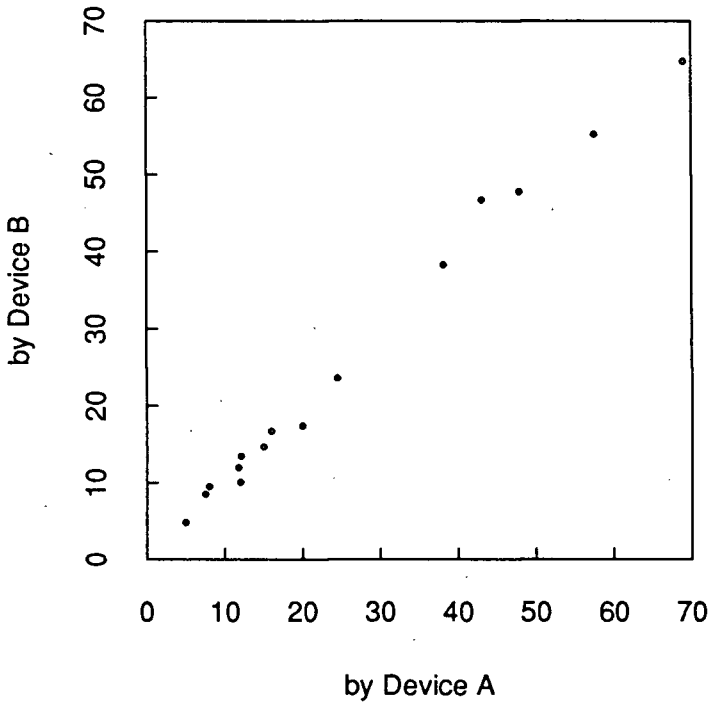


Average-difference plot

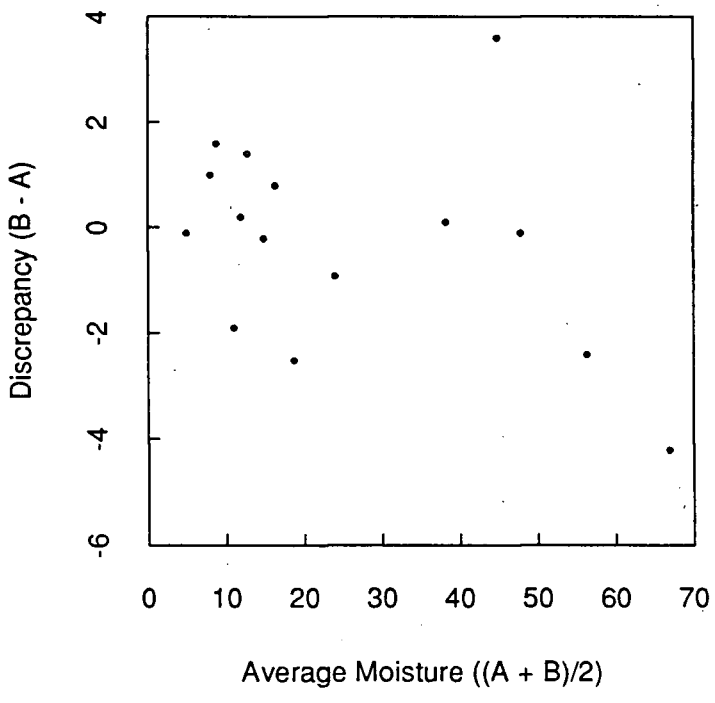


Figures 19 & 20: Tobacco moisture content
(Data Set 6, n = 15)

Scatter plot



Average-difference plot



BIBLIOGRAPHY

[1] Altman, D.G., and Bland, M.J. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician* **32**:307–317.

[2] Amemiya, T. (1973). Regression analysis when the variance of the dependent variable is proportional to the square of its expectation. *Journal of the American Statistical Association* **68**:928–934.

[3] Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Report* **19**:3–11.

[4] Bartlett, M.S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics* **5**:207–212.

[5] Becker, R.A., and Chambers, J.M. (1984). *S: An Interactive Environment for Data Analysis and Graphics.* Belmont: Wadsworth.

[6] Beeler, M.F. (1986). Can we use results of better statistical approaches to method comparison studies? *American Journal of Clinical Pathology* **86**:406.

[7] Bement, T.R., and Williams, J.S. (1969). Variance of weighted regression estimators when sampling errors are independent and heteroscedastic. *Journal of the American Statistical Association* **64**:1369–1382.

[8] Benjamini, Y. (1983). Is the t test really conservative when the parent distribution is long-tailed? *Journal of the American Statistical Association* **78**:645–54.

[9] Bland, M.J., and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* No.8476 (Feb 8):307–310.

[10] Bottos, M., Petterazzo, A., Giancola, G., Stefani, D., Pettená, G., Viscolani, B., and Rubaltelli, F. (1985). The effect of a 'containing' position in a hammock versus the supine position on the cutaneous oxygen level in premature and term babies. *Early Human Development* **11**:265–273.

[11] Cassidy, P.G., Triplett, D.A., and LaDuca, F.M. (1985). Use of the agarose gel method to identify and quantitate Factor VIII:C inhibitors. *American Journal of Clinical Pathology* **83**:697–706.

[12] Cicchetti, D.V. (1976). Assessing inter-rater reliability for rating scales: resolving some basic issues. *British Journal of Psychiatry* **129**:452–456.

[13] Cleveland, W.S. (1985). *The Elements of Graphing Data.* Monterey: Wadsworth.

[14] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**:37–46.

[15] ———. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**:213–220.

[16] Conover, W.J. (1973). On methods of handling ties in the Wilcoxon signed-rank test. *Journal of the American Statistical Association* **68**:985–988.

[17] Cressie, N. (1980). Relaxing assumptions in the one sample t-test. *Australian Journal of Statistics* **22**:143–153.

[18] Cureton, E.E. (1967). The normal approximation to the signed-rank sampling distribution when zero differences are present. *Journal of the American Statistical Association* **62**:1068–1069.

[19] Daniel, W.W. (1983). *Biostatistics: A Foundation for Analysis in the Health Sciences*, 3rd ed. New York: John Wiley & Sons.

[20] Deming, W.E. (1943). *Statistical Adjustment of Data.* New York: John Wiley & Sons.

[21] Draper, N.R., and Smith H. (1981). *Applied Regression Analysis*, 2nd ed. New York: John Wiley & Sons.

[22] Ebel, R.L. (1951). Estimation of the reliability of ratings. *Psychometrika* **16**: 407–424.

[23] Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association* **64**:1278–1302.

[24] ———. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**:1–26.

[25] ———. (1981). Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics* **9**:139–172.

[26] ———. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: SIAM.

[27] ———. (1985). Bootstrap confidence intervals for parametric problems. *Biometrika* **72**:45–58.

[28] Feldman, S., Klein, D.F., and Honingfeld, G. (1972). The reliability of a decision tree technique applied to psychiatric diagnosis. *Biometrics* **28**:831–840.

[29] Fisher R.A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B* **22**:69–78.

[30] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd ed. New York: John Wiley & Sons.

[31] ———. (1986). *The Design and Analysis of Clinical Experiments.* New York: John Wiley & Sons.

[32] ———, and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* **33**:613–619.

[33] Freedman, D., Pisani, R., and Purves, R. (1978). *Statistics.* New York: Norton.

[34] Gayen, A.K. (1949). The distribution of "Student's" t in random samples of any sizes drawn from non-normal universes. *Biometrika* **36**:353–369.

[35] Geary, R.C. (1936). The distribution of "Student's" ratio of non-normal samples. *Journal of the Royal Statistical Society, Series B* **3**:178–184.

[36] ———. (1947). Testing for normality. *Biometrika* **34**:209–242.

[37] Grubbs, F.E. (1948). On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association* **43**:243–264.

[38] Guilford, J.P. (1954). *Psychometric Methods.* New York: McGraw-Hill.

[39] Gulliksen, H. (1950). *Theory of Mental Tests.* New York: Wiley.

[40] Haggard, E.A. (1958). *Intraclass Correlation and the Analysis of Variance.* New York: Dryden.

[41] Harvard (1955). *Tables of the Cumulative Binomial Probability Distribution.* Cambridge, Massachusetts: Harvard University Press.

[42] Hemelrijk ,J. (1952). A theorem on the sign test when ties are present. *Indagationes Mathematicae* **14**:322–326.

[43] Hinkley, D.V. (1976). On estimating a symmetric distribution. *Biometrika* **63**:680.

[44] Hoffman, P.J. (1963). Test reliability and practice effects. *Psychometrika* **28**:273–288.

[45] Hollander, M., and Wolfe D.A. (1973). *Nonparametric Statistical Methods.* New York: John Wiley & Sons.

[46] Hoyt, C.J., and Krishnaiah P.R. (1960). Estimation of test reliability by analysis of variance technique. *Journal of Experimental Education* **28**:257–259.

[47] Jacquez, J.A., Mather, F.J., and Crawford, C.A. (1968). Linear regression with non-constant unknown error variance: sampling experiments with least squares, weighted least squares, and maximum likelihood estimators. *Biometrics* **24**:607–626.

[48] Kotz, S., and Johnson, N. (1982). *Encyclopedia of Statistical Sciences, Vol 2.* New York: John Wiley & Sons.

[49] Krauth, J. (1973). An asymptotic UMP sign test in the presence of ties. *Annals of Statistics* **1**:166–169.

[50] Landis, J.R., and Koch G.G. (1975). A review of statistical methods in analysis of data arising from observer reliability studies (Parts I & II). *Statistica Neerlandica* **29**:101–123 & 151–161.

[51] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

[52] ———. (1983). *Theory of Point Estimation.* New York: John Wiley & Sons.

[53] Madansky, A. (1959). The fitting of straight line when both variables are subject to error. *Journal of the American Statistical Association* **54**:173–205.

[54] Mandel, J. (1964). *The Statistical Analysis of Experimental Data.* New York: Interscience Publishers.

[55] ———. (1984). Fitting straight lines when both variables are subject to error. *Journal of Quality Technology* **16**:1–14.

[56] Noether, G.E. (1967). *Elements of Nonparametric Statistics.* New York: John Wiley & Sons.

[57] Owen, D.B. (1962). *Handbook of Statistical Tables.* Reading, Massachusetts: Addison-Wesley.

[58] Pearson, E.S., and Adyanthaye, N.K. (1929). The distribution of frequency constants in small samples from non-normal symmetric and skew populations. *Biometrika* **21**:259–286.

[59] Pearson, E.S., and Hartley, H.O. (1972). *Biometrika Tables for Statisticians, Vol 2.* Reprint with corrections, 1976. Cambridge: Cambridge University Press.

[60] Pearson, E.S., and Please, N.W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika.* **62**:223–241.

[61] Pitman, E.J.G. (1948). *Lecture Notes on Nonparametric Statistical Inference.* New York: Columbia University Press.

[62] Pratt, J.W. (1959). Remark on zeros and ties in the Wilcoxon signed-rank procedures. *Journal of the American Statistical Association* **54**:655–667.

[63] ———. (1961). Length of confidence intervals. *Journal of the American Statistical Association* **56**:549–567.

[64] ———, and Gibbons, J.D. (1981). *Concepts of Nonparametric Theory.* New York: Springer-Verlag.

[65] Putter, J. (1955). The treatment of ties in some nonparametric tests. *Annals of Mathematical Statistics* **26**:368–386.

[66] Rahe, A.J. (1974). Tables of critical values for the Pratt matched pair signed rank statistic. *Journal of the American Statistical Association* **69**:368–373.

[67] Randles, R.H., and Hogg, R.V. (1973). Adaptive distribution-free tests. *Communications in Statistics* **2**:337–356.

[68] Randles, R.H., and Wolfe, D.A. (1979). *Introduction of the Theory of Nonparametric Statistics.* New York: John Wiley & Sons.

[69] Rawles, J. (1986). Regression analysis. *The Lancet* No.8481 (March 15):614.

[70] Robinson, E.F., and Wade, W.D. (1983). Statistical assessment of two methods of measuring scoliosis before treatment. *Canadian Medical Association Journal* **21**:839–841.

[71] Rubin, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics* **9**:130–134.

[72] Ryan, T.A.Jr., Joiner, B.L., and Ryan, B.F. (1985). *Minitab Student Handbook*, 2nd ed. Boston: Duxbury Press.

[73] Sarhan, A.E., and Greenberg, B.G. (1962). *Contributions to Order Statistics.* New York: John Wiley & Sons.

[74] Thompson, W.A.Jr. (1963). Precision of simultaneous measurement procedures. *Journal of the American Statistical Association* **58**:474–479.

[75] Tibshirani, R.J. (1984). Bootstrap confidence intervals. Stanford University, Division of Biostatistics Technical Report No. **91**.

[76] Winer, B.J. (1962). *Statistical Principles in Experimental Design.* New York: McGraw-Hill.

APPENDIX I. BOOTSTRAP METHOD

1. GENERAL THEORY OF BOOTSTRAP METHODS

Suppose we wish to draw inferences about some parameter θ of a population with unknown distribution F based on realization of an independent identically distributed sample $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ from F. It may be convenient to denote the parameter of interest by $\theta(F)$. And suppose $\hat{\theta}$ is an estimator of θ; we also write $\hat{\theta}$ as $\hat{\theta}(X_1, X_2, \cdots, X_n)$ to indicate that the statistic is a function of $X_1, X_2, \cdots, X_n$.

Let $\hat{F}$ be the empirical distribution of the random sample, putting probability mass $\frac{1}{n}$ on each x_i; and let $X_1^*, X_2^*, \cdots, X_n^*$ denote a random sample from $\hat{F}$, i.e., drawn independently *with replacement* from $\{x_1, x_2, \cdots, x_n\}$:

$$X_1^*, X_2^*, \cdots, X_n^* \overset{\text{iid}}{\sim} \hat{F}.$$

Call $X_1^*, X_2^*, \cdots, X_n^*$ a "bootstrap sample". Then $\hat{\theta}^* \equiv \hat{\theta}(X_1^*, X_2^*, \cdots, X_n^*)$ estimates $\theta(\hat{F})$, considering $\hat{F}$ as fixed, that is, conditioning on the sample values.

In theory, inferences about the parameter $\theta(F)$ can be based on the distribution of $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_1)$; and behaviour of $\hat{\theta}$ can be approximated

by behaviour of $\hat{\theta}^* = \hat{\theta}(X_1^*, X_2^*, \cdots, X_n^*)$. The distribution of $\hat{\theta}^*$, in general, may be difficult to obtain analytically; but it can always be approximated by using a Monte Carlo algorithm, as discussed below.

Suppose we know that the probability distribution F is symmetric. In this case, we would symmetrize $\hat{F}$. One way to achieve this is to replace $\hat{F}$ by $\hat{F}_{SYM}$, the symmetric probability distribution obtained from $\hat{F}$ by reflection about the median. That is, $\hat{F}_{SYM}$ has probability mass $\frac{1}{2n-1}$ on each $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$, and $2x_{(m)} - x_{(1)}, 2x_{(m)} - x_{(2)}, \cdots, 2x_{(m)} - x_{(n)}$, assuming that n is odd and equal to $2m-1$ for convenience [Efron (1979)]. In this case, even though the symmetrized distribution is not a nonparametric maximum likelihood estimate for F, the symmetrized distribution has properties similar to the maximum likelihood estimate, $\hat{F}$ [Hinkley (1976)].

2. BOOTSTRAP ESTIMATOR AND CONFIDENCE INTERVAL

Recall that bootstrap estimators or confidence intervals for θ rely on the distribution of $\hat{\theta}^* = \hat{\theta}(X_1^*, X_2^*, \cdots, X_n^*)$ — the estimator $\hat{\theta}$ evaluated at a "bootstrap sample" $\{X_1^*, X_2^*, \cdots, X_n^*\}$ generated as independent and identically distributed observations from the empirical distribution $\hat{F}$. As already noted, in general, the distribution of $\hat{\theta}^*$ is hard to find analytically, but can be

approximated by a Monte Carlo method. But for the sample mean and the sample median, the bootstrap distribution can be obtained theoretically, without using the Monte Carlo methods.

2.1. BOOTSTRAPPING FOR THE MEAN

For the mean, the parameter of interest is $\theta(F) = E(X)$. So, $\hat{\theta} = \bar{X}$, the sample mean, is the estimator of $\theta(F)$.

It can be shown that

$$E_*(\bar{X}^*) = \bar{X}, \quad \text{and}$$

$$Var_*(\bar{X}^*) = \frac{1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\hat{\sigma}^2.$$

Also, the central limit theorem implies that the bootstrap distribution of $\bar{X}^*$ is approximately normal, $\mathcal{N}(\bar{X}, \frac{1}{n}\hat{\sigma}^2)$. Thus, by the central limit theorem, the bootstrap interval estimate would essentially be the same as the t interval estimate as derived in Section 3.1.

2.2. BOOTSTRAPPING FOR THE MEDIAN

For the median, the parameter of interest is the point θ such that

$$Prob(X < \theta) \leq \frac{1}{2} \leq Prob(X \leq \theta).$$

So, the estimator is $\hat{\theta} = \tilde{X}$, the sample median.

Suppose $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ is the realization of a sample. For convenience, suppose the sample size is odd and equal to $2m-1$, say. Then the sample median estimate of $\theta(F)$ is $\tilde{X} = x_{(m)}$. Then the bootstrap distribution of $\hat{\theta}^*$ is concentrated on the values $x_{(1)} \leq x_{(2)} \leq \cdots \leq, x_{(n)}$ such that

$$p_k \equiv Prob_* \left[\hat{\theta}_* = x_{(k)}\right]$$
$$= \sum_{i=1}^{m-1} \left\{ \binom{n}{k-1} \left(\frac{k-1}{n}\right)^j \left(\frac{n-k-1}{n}\right)^{n-j} - \binom{n}{k} \left(\frac{k}{n}\right)^j \left(\frac{n-k}{n}\right)^{n-j} \right\}$$

[Efron (1982, p.77)].

Furthermore, Efron showed that the corresponding confidence interval is very close to the classical interval estimate for the median as discussed in Section 3.2 [Efron (1982, pp.80-81)].

2.3. MONTE CARLO EVALUATION OF THE BOOTSTRAP DISTRIBUTION FOR ARBITRARY $\hat{\theta}^*$

1. Construct the nonparametric maximum likelihood estimator of F, the empirical distribution $\hat{F}$,

$$\hat{F}: \text{ mass } \frac{1}{n} \text{ at } x_1, x_2, \cdots, x_n.$$

(For emphasis, we could write $F_X \equiv F$ and $\hat{F}_X \equiv \hat{F}$). In the symmetric bootstrap case, replace $\hat{F}$ by $\hat{F}_{SYM}$,

$\hat{F}_{SYM}$: mass $\frac{1}{2n-1}$ at $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$, and

$$2x_{(m)} - x_{(1)}, 2x_{(m)} - x_{(2)}, \cdots, 2x_{(m)} - x_{(n)},$$

assuming that n is odd and equal to $2m-1$, say, for convenience [Efron (1979)].

2. Draw a "bootstrap sample" of size n from $\hat{F}$,

$$X_1^*, X_2^*, \cdots, X_n^* \overset{\text{iid}}{\sim} \hat{F}$$

and calcalute $\hat{\theta}^* = \hat{\theta}(X_1^*, X_2^*, \cdots, X_n^*)$.

3. Independently repeat step 2 B times (for some large B), obtaining "bootstrap replications" $\left\{\hat{\theta}_b^*,\ b = 1, 2, \cdots, B\right\}$.

4. Approximate the cumulative distribution function of $\hat{\theta}^*$ by the empirical cumulative distribution function of $\left\{\hat{\theta}_b^*,\ b = 1, 2, \cdots, B\right\}$:

$$\begin{aligned} \hat{F}_{\hat{\theta}^*}(t) &= \frac{1}{B}\{\#\text{ of } \hat{\theta}_b^* \le t\} \\ &= \frac{1}{B}\sum_{b=1}^{B} I\{\hat{\theta}_b^* \le t\} \ , \end{aligned}$$

where the indicator function

$$I\{\hat{\theta}_b^* \le t\} = \begin{cases} 1, & \text{if } \hat{\theta}_b^* \le t; \\ 0, & \text{otherwise.} \end{cases}$$

[Efron (1982, p.28)].

2.4. BIAS-CORRECTED BOOTSTRAP ESTIMATE

Statistic $\hat{\theta}$ need not be unbiased; in general

$$\text{Bias} = E_F\hat{\theta} - \theta,$$

where E_F indicates expectation is taken with respect to the distribution F. This bias can be approximated by

$$\text{Bias}^* = E_*\hat{\theta}^* - \hat{\theta};$$

where E_* denotes expectation with respect to $\hat{F}$; and a Monte Carlo approximation of Bias* is given by

$$\begin{aligned}\widehat{\text{Bias}^*} &= \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^* - \hat{\theta} \\ &= \bar{\theta}^* - \hat{\theta},\end{aligned}$$

where $\bar{\theta}^*$ is the average of the B bootstrap replications of $\hat{\theta}^*, \left\{\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*\right\}$

Thus, a bias-corrected bootstrap estimate is given by

$$\begin{aligned}\hat{\theta}_B &= \hat{\theta} - (\bar{\theta}^* - \hat{\theta}) \\ &= 2\hat{\theta} - \bar{\theta}^*.\end{aligned}$$

However, this bias-corrected estimate would have a larger variance than the original estimate $\hat{\theta}$ because

$$Var(\hat{\theta}_B) = Var(\hat{\theta} - \widehat{Bias}^*)$$

$$= Var(\hat{\theta}) + Var(\widehat{Bias}^*) - 2Cov(\hat{\theta}, \widehat{Bias}^*),$$

where usually $2Cov(\hat{\theta}, \widehat{Bias}^*) \approx 0$. (This is a "variance-bias tradeoff").

And the bootstrap estimate of standard error of $\hat{\theta}$ is equal to $\sigma^*(\hat{\theta})$, which can be estimated by sample standard deviation of the bootstrap replication of $\hat{\theta}^*$:

$$\hat{\sigma}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \bar{\theta}^*\right)^2}$$

[Efron (1982, p.28)].

2.5. BOOTSTRAP CONFIDENCE INTERVALS

Let $F_{\hat{\theta}^*}(t) = Prob_*\{\hat{\theta}^* \leq t\}$ be the cumulative distribution function of $\hat{\theta}^*$. Note that if the bootstrap distribution is obtained by the Monte Carlo methods, then $F_{\hat{\theta}^*}(t)$ is approximated by the empirical cumulative distribution function of the bootstrap replications of $\hat{\theta}^*$, $\left\{\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*\right\}$:

$$F_{\hat{\theta}^*}(t) \approx \frac{1}{B}\{\#\ \text{of}\ \hat{\theta}_b^* \leq t\} = \frac{1}{B}\sum_{b=1}^{B} I\{\hat{\theta}_b^* \leq t\},$$

where the indicator function

$$I\{\hat{\theta}_b^* \leq t\} = \begin{cases} 1, & \text{if } \hat{\theta}_b^* \leq t; \\ 0, & \text{otherwise.} \end{cases}$$

2.5.1. THE PERCENTILE METHOD

Suppose that $\hat{\theta} - \theta$ is a pivotal quantity, that is

$$\hat{\theta} - \theta \sim H, \qquad (A1)$$

where H is a distribution *not* involving θ. Also, suppose that approximately

$$\hat{\theta}^* - \hat{\theta} \overset{*}{\sim} H. \qquad (A2)$$

In (A_2) the distribution $\hat{F}$ plays the same role as F in (A_1), "$\overset{*}{\sim}$" indicating the distribution under independent and identical sampling from $\hat{F}$. Notice that the second assumption is reasonable because, if $\hat{F}$ is close to F, then the bootstrap distribution of $\hat{\theta}^* - \hat{\theta}$ will be "close" to that of $\hat{\theta} - \theta$, as long as $\theta(\cdot)$ is a reasonably smooth functional. Finally, assume that H is symmetric about 0. $(A3)$

Then a $1 - 2\alpha$ symmetric confidence interval $(\theta_{LOW}, \theta_{UP})$ is given by

$$\theta_{LOW} = F^{-1}_{\hat{\theta}^*}(\alpha) \quad \text{and} \quad \theta_{UP} = F^{-1}_{\hat{\theta}^*}(1 - \alpha)$$

[Efron (1982, p.78)].

Assumptions (A1) and (A2) can be generalized to

$$g(\hat{\theta}) - g(\theta) \sim H, \qquad (A1')$$

and

$$g(\hat{\theta}^*) - g(\hat{\theta}) \overset{*}{\sim} H, \qquad (A2')$$

where H symmetric about 0 and $g(\cdot)$ is an unknown, monotone increasing function. Indeed, further knowledge about $g(\cdot)$ is not necessary since the resultant interval does not depend on $g(\cdot)$ [Tibshirani (1984)]. These procedures simply assume the existence of a symmetric pivotal on some other scale.

Under these generalized assumptions, the interval $\left(F^{-1}_{\hat{\theta}^*}(\alpha), F^{-1}_{\hat{\theta}^*}(1-\alpha)\right)$ remains valid as a $1-2\alpha$ confidence interval [Tibshirani (1984)].

2.5.2. THE BIAS-CORRECTED PERCENTILE METHOD

If H, the distribution of the pivotal quantity $g(\hat{\theta}) - g(\theta)$, is symmetric about a point, say μ, which does not equal 0, then the percentile interval will be biased and will not have the correct coverage. In order to estimate μ and hence to derive a bias correction to the percentile interval, we need to assume a parametric form for H. Tibshirani (1984) showed that the bias-corrected percentile interval is robust with respect to the choice of a symmetric pivotal distribution.

Suppose $H = \mathcal{N}(\mu, 1)$. Define

$$z_0 = \Phi^{-1}\left(F_{\hat{\theta}^*}(\hat{\theta})\right)$$

to estimate μ, where Φ is the cumulative distribution function of $\mathcal{N}(0,1)$. Then a $1-2\alpha$ bias-corrected percentile interval $(\theta_{LOW}, \theta_{UP})$ is given by

$$\theta_{LOW} = F^{-1}_{\hat{\theta}^*}\left(\Phi(2z_0 - z_\alpha)\right) \quad \text{and} \quad \theta_{UP} = F^{-1}_{\hat{\theta}^*}\left(\Phi(2z_0 + z_\alpha)\right)$$

[Efron (1982, p.82)].

Efron (1982, p.86) remarked that the bias-corrected percentile interval should be used with caution, or not at all, when distributional asymmetry is definite.

APPENDIX II. EFFICACY CALCULATIONS AND ARE'S FOR STANDARD DISTRIBUTIONS AND MIXTURES

Recall that for two statistical tests, say T_1 and T_2, Pitman asymptotic relative efficiency (ARE) of T_1 with respect to T_2 can be represented as a squared ratio of efficacies:

$$ARE(T_1, T_2) = \left[\frac{eff(T_1)}{eff(T_2)}\right]^2,$$

where $eff(T_i)$ denotes the efficacy of test T_i [Randles and Wolfe (1979, pp.147-149)].

Suppose K is a statistic used by test T for a hypothesis $\theta = \theta_0$, where θ is a parameter of a symmetric density. Suppose we reject the hypothesis if K is outside a certain interval. Then the efficacy of T, denoted by $eff(T)$, is defined as

$$eff(T) = \frac{\left[\frac{\partial E_\theta(K)}{\partial \theta}\right]^2\Big|_{\theta=\theta_0}}{Var_{\theta_0}(K)}$$

[Kotz and Johnson (1982, p.468)].

For a density function f_X symmetric about 0, the efficacies of t test

(T), sign test (S), and Wilcoxon signed-rank test (W) are given by

$$eff(T) = \frac{1}{\sigma_X},$$
$$eff(S) = 2f_X(0), \quad \text{and}$$
$$eff(W) = 2\sqrt{3}\int_{-\infty}^{\infty} f_X^2(x)dx,$$

where $\sigma_X^2 = Var(X) = \int_{-\infty}^{\infty} x^2 f_X(x)dx$ [Randles and Wolfe (1979, pp.165-168)].

These efficacies are well known for particular symmetric families, including normal, uniform, and Cauchy densities. Note that although efficacy may be a function of family parameter(s) (normal standard deviation σ, etc.), efficacy *ratios* — that is, Pitman relative efficiencies — are not. Hence, it suffices to evaluate numerically the efficacy of standard normal density, etc., as in Table 1, adapted from Pratt and Gibbons (1981, p.384).

Efficacy and ARE also can be calculated for mixtures of normal distributions. Suppose $X \sim \sum_{i=1}^{k} \omega_i X_i$, where $X_i \overset{\text{indep}}{\sim} \mathcal{N}(0, \sigma_i^2)$, and ω_i are weights such that $0 \leq \omega_i \leq 1$, $\sum_{i=1}^{k} \omega_1 = 1$. Then $f_X(x) = \sum_{i=1}^{k} \omega_i \frac{1}{\sqrt{2\pi}\sigma_i} exp\left(\frac{-x^2}{2\sigma_i^2}\right)$.

We need to compute σ_X^2, $f_X(0)$, and $\int_{-\infty}^{\infty} f_X^2(x)dx$ in order to obtain

the efficacies of T, S, and W for the mixture.

$$\begin{aligned}\sigma_X^2 &= \int_{-\infty}^{\infty} x^2 \sum_{i=1}^{k} \frac{\omega_i}{\sqrt{2\pi}\sigma_i} exp\left(\frac{-x^2}{2\sigma_i^2}\right) dx \\ &= \sum_{i=1}^{k} \omega_i \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}\sigma_i} exp\left(\frac{-x^2}{2\sigma_i^2}\right) dx \\ &= \sum_{i=1}^{k} \omega_i \sigma_i^2.\end{aligned}$$

$$f_X(0) = \sum_{i=1}^{k} \frac{\omega_i}{\sqrt{2\pi}\sigma_i} exp(0) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{k} \frac{\omega_i}{\sigma_i}.$$

$$\begin{aligned}\int_{-\infty}^{\infty} f_X^2(x)dx &= \int_{-\infty}^{\infty} \left\{ \sum_{i=1}^{k} \frac{\omega_i}{\sqrt{2\pi}\sigma_i} exp\left(\frac{-x^2}{2\sigma_i^2}\right) \right\}^2 dx \\ &= \int_{-\infty}^{\infty} \left\{ \sum_{i=1}^{k} \frac{\omega_i^2}{2\pi\sigma_i^2} exp\left(\frac{-x^2}{\sigma_i^2}\right) + \sum_{i=1}^{k} \sum_{j<i} \frac{\omega_i\omega_j}{2\pi\sigma_i\sigma_j} exp\left[\frac{-x^2}{2}\left(1/\sigma_i^2 + 1/\sigma_j^2\right)\right] \right\} dx \\ &= \frac{1}{\sqrt{2\pi}} \left\{ \sum_{i=1}^{k} \frac{\omega_i^2}{\sqrt{2}\sigma_i} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}(\sigma_i/\sqrt{2})} exp\left[\frac{-x^2}{2(\sigma_i^2/2)}\right] dx \; + \right. \\ &\quad \left. 2 \sum_{i=1}^{k} \sum_{j<i} \frac{\omega_i\omega_j}{\sqrt{\sigma_i^2+\sigma_j^2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\left(\sigma_i\sigma_j/\sqrt{\sigma_i^2+\sigma_j^2}\right)} exp\left[\frac{-x^2}{2\left(\sigma_i^2\sigma_j^2/(\sigma_i^2+\sigma_j^2)\right)}\right] dx \right\} \\ &= \frac{1}{\sqrt{2\pi}} \left\{ \sum_{i=1}^{k} \frac{\omega_i^2}{\sqrt{2}\sigma_i} + 2 \sum_{i=1}^{k} \sum_{j<i} \frac{\omega_i\omega_j}{\sqrt{\sigma_i^2+\sigma_j^2}} \right\}\end{aligned}$$

since both integrands are density functions of normals.

EXAMPLE 1: Consider a mixture with equal proportions of four normals,

$\mathcal{N}(0, i^2), i = 1, 2, 3, 4$. That is, $\omega_i = \frac{1}{4}$ and $\sigma_i = i$, for $i = 1, 2, 3, 4$. Then

$$\sigma_X^2 = \sum_{i=1}^{4} \omega_i \sigma_i^2 = \frac{1}{4} \sum_{i=1}^{4} i^2 = \frac{15}{2},$$

$$f_X(0) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{4} \frac{\omega_i}{\sigma_i} = \frac{1}{4\sqrt{2\pi}} \sum_{i=1}^{4} \frac{1}{i} = \frac{1}{4\sqrt{2\pi}} \left(\frac{25}{12}\right) = \frac{25}{48\sqrt{2\pi}}, \quad \text{and}$$

$$\begin{aligned} \int_{-\infty}^{\infty} f_X^2(x)dx &= \frac{1}{\sqrt{2\pi}} \left\{ \sum_{i=1}^{4} \frac{\omega_i^2}{\sqrt{2}\sigma_i} + 2 \sum_{i=1}^{4} \sum_{j<i} \frac{\omega_i \omega_j}{\sqrt{\sigma_i^2 + \sigma_j^2}} \right\} \\ &= \frac{1}{16\sqrt{2\pi}} \left\{ \sum_{i=1}^{4} \frac{1}{\sqrt{2}i} + 2 \sum_{i=1}^{4} \sum_{j<i} \frac{1}{\sqrt{i^2 + j^2}} \right\} \\ &= \frac{1}{16\sqrt{2\pi}} (4.8870). \end{aligned}$$

Thus,

$$ARE(W, S) = 3 \left[\frac{\int_{-\infty}^{\infty} f_X^2(x)dx}{f_X(0)} \right]^2 = 3 \left[\frac{3(4.8870)}{25} \right]^2 = 1.0317,$$

$$ARE(S, T) = 4\sigma_X^2 f_X^2(0) = 4 \left(\frac{15}{2}\right) \left(\frac{25}{48\sqrt{2\pi}}\right)^2 = 1.2952, \quad \text{and}$$

$$ARE(W, T) = 12\sigma_X^2 \left[\int_{-\infty}^{\infty} f_X^2(x)dx \right]^2 = 12 \left(\frac{15}{2}\right) \left[\frac{1}{16\sqrt{2\pi}} (4.8870) \right]^2 = 1.3363.$$

Recall that, for corresponding confidence intervals, the asymptotic ratio of interval lengths $\frac{L_1}{L_2} = \frac{1}{\sqrt{ARE(1,2)}}$; and note that the ratios all are close to one for the mixture in this example.

EXAMPLE 2: Consider a mixture with equal proportions of five normals,

$\mathcal{N}(0, i^2), i = 1, 2, 3, 4, 5$. That is, $\omega_i = \frac{1}{5}$ and $\sigma_i = i$, for $i = 1, 2, 3, 4, 5$. Then

$$\sigma_X^2 = \sum_{i=1}^{5} \omega_i \sigma_i^2 = \frac{1}{5} \sum_{i=1}^{5} i^2 = 11,$$

$$f_X(0) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{5} \frac{\omega_i}{\sigma_i} = \frac{1}{5\sqrt{2\pi}} \sum_{i=1}^{5} \frac{1}{i} = \frac{1}{5\sqrt{2\pi}} \left(\frac{137}{60} \right), \quad \text{and}$$

$$\begin{aligned} \int_{-\infty}^{\infty} f_X^2(x) dx &= \frac{1}{\sqrt{2\pi}} \left\{ \sum_{i=1}^{5} \frac{\omega_i^2}{\sqrt{2}\sigma_i} + 2 \sum_{i=1}^{5} \sum_{j<i} \frac{\omega_i \omega_j}{\sqrt{\sigma_i^2 + \sigma_j^2}} \right\} \\ &= \frac{1}{25\sqrt{2\pi}} \left\{ \sum_{i=1}^{5} \frac{1}{\sqrt{2}i} + 2 \sum_{i=1}^{5} \sum_{j<i} \frac{1}{\sqrt{i^2 + j^2}} \right\} \\ &= \frac{1}{25\sqrt{2\pi}} (6.4474). \end{aligned}$$

Thus,

$$ARE(W, S) = 3 \left[\frac{\int_{-\infty}^{\infty} f_X^2(x) dx}{f_X(0)} \right]^2 = 3 \left[\frac{60(6.4474)}{5(137)} \right]^2 = 0.9568,$$

$$ARE(S, T) = 4\sigma_X^2 f_X^2(0) = 4(11) \left[\frac{1}{5\sqrt{2\pi}} \left(\frac{137}{60} \right) \right]^2 = 1.4604, \quad \text{and}$$

$$ARE(W, T) = 12\sigma_X^2 \left[\int_{-\infty}^{\infty} f_X^2(x) dx \right]^2 = 12(11) \left[\frac{1}{25\sqrt{2\pi}} (6.4474) \right]^2 = 1.3973.$$

Note that, although this example is very similar to the preceding mixture, numeric results for $ARE(W, S)$ are less than one here.

EXAMPLE 3: Consider a standard normal $\mathcal{N}(0, 1)$ contaminated with 5%

of $\mathcal{N}(0, 10^2)$. That is, $\omega_1 = 0.95$, $\omega_2 = 0.05$, $\sigma_1 = 1$, and $\sigma_2 = 10$. Then

$$\sigma_X^2 = \sum_{i=1}^{2} \omega_i \sigma_i^2 = 0.95(1) + 0.05(100) = 1.45,$$

$$f_X(0) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{2} \frac{\omega_i}{\sigma_i} = \frac{1}{\sqrt{2\pi}} \left\{ 0.95 + \frac{0.05}{10} \right\} = \frac{0.955}{\sqrt{2\pi}}, \quad \text{and}$$

$$\begin{aligned} \int_{-\infty}^{\infty} f_X^2(x)dx &= \frac{1}{\sqrt{2\pi}} \left\{ \sum_{i=1}^{2} \frac{\omega_i^2}{\sqrt{2}\sigma_i} + 2 \sum_{i=1}^{2} \sum_{j<i} \frac{\omega_i \omega_j}{\sqrt{\sigma_i^2 + \sigma_j^2}} \right\} \\ &= \frac{1}{\sqrt{2\pi}} \left\{ \frac{(0.95)^2}{\sqrt{2}} + \frac{(0.05)^2}{10\sqrt{2}} + \frac{2(0.95)(0.05)}{\sqrt{1+100}} \right\} \\ &= \frac{1}{\sqrt{2\pi}} (0.6478). \end{aligned}$$

Thus,

$$ARE(W, S) = 3 \left[\frac{\int_{-\infty}^{\infty} f_X^2(x)dx}{f_X(0)} \right]^2 = 3 \left[\frac{0.6478}{0.955} \right]^2 = 1.3804,$$

$$ARE(S, T) = 4\sigma_X^2 f_X^2(0) = 4(1.45) \left[\frac{0.955}{\sqrt{2\pi}} \right]^2 = 0.8419, \quad \text{and}$$

$$ARE(W, T) = 12\sigma_X^2 \left[\int_{-\infty}^{\infty} f_X^2(x)dx \right]^2 = 12(1.45) \left[\frac{0.6478}{\sqrt{2\pi}} \right]^2 = 1.1621.$$

Note that, for this contaminated normal, W becomes preferable to T, contrasting with the result for pure normal.

APPENDIX III TAIL-WEIGHT ADAPTIVE NONPARAMETRIC PROCEDURES

Given a sample of discrepancies $D_1, D_2, \cdots, D_n$, Randles and Hogg (1973) defined a tail-weight statistic

$$Q = \frac{10(U_{0.05} - L_{0.05})}{U_{0.50} - L_{0.50}},$$

where U_β (L_β) is the sum of the largest (smallest) $n\beta$ order statistics (fractional items are used if $n\beta$ is not an integer). For instance, if $n = 26$, $n(0.05) = 1.3$; so $L_{0.05} = D_{(26)} + 0.3D_{(25)}$.

Then the underlying distribution will be classified as having light, moderate, or heavy tails if $Q < 2.08 - \frac{2}{n}$, $2.08 - \frac{2}{n} \leq Q \leq 2.96 - \frac{5.5}{n}$, or $Q > 2.96 - \frac{5.5}{n}$, respectively.

Randles and Hogg (1973) showed that the Q statistic discussed above is uncorrelated with the Student t, sign, and Wilcoxon signed rank statistics. A zero correlation would give asymptotic independence, because the relevant joint distribution is asymptotically normal; but, independence does not hold for finite samples. In particular, simulation gives confidence intervals with confidence less than the nominal level for $n = 18$; see Randles and Hogg (1973).

To obtain a truly nonparametric adaptive procedure, Randles and Hogg modified the tail-weight statistic Q as

$$Q^* = \frac{100U^*_{0.10}}{\sum_{i=1}^{n} |D_i|},$$

where $U^*_{0.10}$ is the sum of the largest 10% values $|D_1|, |D_2|, \cdots, |D_n|$, and the classification rule works as above. Notice that Q^* is independent of all rank statistics, like the Student t (viewed as an approximation to permutation statistic), sign, and Wilcoxon signed rank statistics, because Q^* is a function of the order statistics of $|D_1|, |D_2|, \cdots, |D_n|$, which are sufficient and complete for a continuous symmetric distribution and because sufficient statistics are independent of every rank statistic [Lehmann (1983, p.40 and p.68)].