39 A SURVEY OF SAMPLING FROM CONTAMINATED DISTRIBUTIONS[1]

JOHN W. TUKEY
Princeton University

1. Historical Introduction

It is particularly fitting that the results produced to date on sampling from contaminated distributions should be summarized in a volume dedicated to Harold Hotelling. Toward the end of World War II, several groups under contract to the Applied Mathematics Panel, National Defense Research Council, had common concern with problems of the effectiveness of bomber machine-gun fire against attacking fighters. Two of these were the Princeton University Fire Control Research Group, of which the writer was a member, and SRG–C (Statistical Research Group–Columbia) for which Harold Hotelling was official investigator. The work summarized in this paper had its origin in those particular problems of accuracy of fire, and, in particular, in the statistical problems of inferring effectiveness from an observed pattern of shots fired with fixed aim or fixed intended aim.

Two contributions to the effectiveness of fire are easily identified: (i) Individual projectiles do not go in the direction in which they are are aimed, or intended to be aimed; (ii) the aim is neither perfect nor constant, and the vulnerability of the target is neither uniform nor concentrated at a point. As a consequence, the probable effectiveness of a projectile traveling at a given angle to the direction of aim, or intended aim, averaged over all situations, decreases smoothly as the angle between the directions of travel and aim increases. Thus, the smaller the dispersion of directions of motion around the direction of aim, the greater probable effectiveness of fire. (Aim-wander, and the possibility of overhitting, cf., for example, Cunningham and Hynd [1], Fraser [2], reverse this trend for sufficiently small dispersions, but the dispersions in question did not fall in this range.)

In assessing records of dispersion, then, the purpose was to assess and compare the average values of such smooth functions. (Because the angular

[1] The basic results surveyed here were obtained in the course of research sponsored by the Office of Naval Research. Their integration and exposition was carried out under Contract No. DA-36-034-ORD-2297 sponsored by the Office of Ordnance Research, U. S. Army.

deviations are small, deviations in a plane perpendicular to the direction of aim or intended aim can replace angles. This is particularly convenient when holes in a flat target produce such deviations as raw data.)

The problem can be further simplified without doing violence to its essentials by treating as known the location of the distribution, as it might be expressed by the coordinate means. Two quite distinct arguments support this simplification. On the one hand, experience with large-sample location-and-scaling problems shows that they behave in very much the same way as the corresponding pure scaling problems. In a sense the location part of the problem is either separable from, or simpler than, the scaling part. On the other hand, it is quite appropriate to believe that, in due course, the adjustment of sighting devices will, in each instance, become accurate enough to place the center of the distribution at the direction of aim, or intended aim. While either argument would probably suffice, together they make the simplification very reasonable.

Thus the assessment-of-dispersion-for-the-purpose-of-assessing-or-comparing-effectiveness problem reduces to a two-dimensional analog of a familiar problem: Given a sample, to assess the scale of the distribution from which it was drawn. If the distribution is normal, then the classical statistical results tell us that we do best to use the three second-degree sample moments. But what if the distribution is not normal?

Some years of close contact with the late C. P. Winsor had taught the writer to beware of extreme deviates, and in particular to beware of using

THE FIRST QUESTION

Given two normal populations with the same mean, one having three times the standard deviation of the other, it is proposed to prepare a sequence of mixed populations by adding varying small amounts of the wider normal population to the narrower one. It is well known that, in large samples, the relative efficiency as a measure of scale of the mean deviation compared with the standard deviation is 88% when the underlying population is normal. As specific small amounts of the wider normal population are added to the narrower one, thus defining new classes of distributions of fixed shape, will the relative efficiency for scaling of the mean deviation compared to the standard deviation increase or decrease?

NOTES: 1. The possible answers are "increase," "stay the same," and "decrease."
2. The problem is a large-sample problem.
3. To find the answer, after giving the question long and careful thought, turn over two pages.

them with high weights. Using second moments to assess variability means giving very high weights to extremely deviant observations. Thus the use of second moments, unquestionably optimum for normal distributions, comes into serious question. When this point was raised in conversation, real differences of opinion between some of the statisticians concerned showed themselves. The earliest work on sampling from contaminated distributions was carried out in an attempt to develop facts which would help to quiet this clash of opinion. (See references [3] through [10] for some of the work done on contaminated distributions.)

2. The Broad Nature of the Problem

Most of the standard statistical techniques are optimum, or recommended, or perhaps merely correctly calibrated, given certain standard assumptions. How do they behave under alternative assumptions? The work to be summarized here can be recognized as another step toward the solution of this broad problem.

It is interesting to speculate why more examples of this broad problem have not been treated. Each specific example can be formulated, and, with the expenditure of sufficient effort, solved within the scope of classical mathematical statistics. But statisticians seem to have been reluctant to approach problems of this type. Is it because the statement of such problems implies that a standard statistical technique might be used when its standard assumptions are not fulfilled? Perhaps a desire to "sweep the dirt under the rug" is responsible for much of the seeming reluctance to tackle such problems. Yet this cannot be the whole explanation, for there have been a fair number of inquiries into the behavior of tests of significance (and, consequently, of techniques of confidence estimation) when something other than the standard assumption holds (e.g., Pearson [11]; Eden and Yates [12]; Pitman [13], [14], [15]; Welch [16]; Box and Andersen [17]). But there has not been any comparable amount of work on questions of the relative precision or accuracy of different point estimates, including questions of relative efficiency. Is this (i) because statisticians do not mean their statements of relative efficiency to be taken seriously, (ii) because statisticians believe that, in contrast to statements of significance, statements of relative efficiency will not be compared with experience, or (iii) because of some reason or reasons not yet suggested? No definitive answer to this question will be presented here, but statisticians owe the problem some consideration.

It would seem that questions of robustness of efficiency are intrinsically at least as important as questions of robustness of significance level. And as soon as it develops, as it shortly will, that failures of robustness of efficiency may be very substantial, the need for more work on the robustness of efficiency will be clear and pressing.

3. Sharpening and Reduction of the Problem

With the problem reduced to a distribution in a plane, a mathematical statistician's first instinct is to consider the special case where this distribution is bivariate normal. This special case is simple, since the first and second moments of the deviations become sufficient statistics for the five parameters of the general case, and, with location assumed known, only second moments need be considered. Any average over the distribution, including the average effectiveness whose assessment is the purpose of the study, is a function of these five parameters and can be efficiently estimated on the basis of (i.e., as a function of) these first and second moments. All this results if the assumption of normality is either to be taken seriously, or to be utilized as a fully effective approximation.

The smooth function of deviation of shot from aim, whose average we seek to assess, needs now to be specified in more detail. The choice

$$\phi = e^{-Q},$$

where Q is a quadratic function of the deviations, is likely to represent the true situation with sufficient accuracy for practical purposes, in part because of our limited knowledge, and in part because great precision is not required.

It is very natural to ask, perhaps in a loud and bitter tone, why this assumed form for the measure of effectiveness should be adequate, when the same assumed form for the probability density (the assumption of normality) is thought to be quite inadequate. There is a simple answer. In dealing with the mathematical statistics of the assumed distribution, there is a temptation to give a very high weight to a very aberrant deviation. If this occurs, very fine detail in the tails of the distribution can matter greatly. But a very aberrant deviation will never be assigned a large probability or a large effectiveness, so that details of the tails of the expression to be averaged will never matter greatly. The artificialities of "efficient" estimation emphasize the tails in a way in which the realities of probability assessment never do.

There is nothing in the problem, as now formulated, which suggests that the two-dimensionality of the situation is playing an essential role. In order to separate essentials more clearly from details, it is thus natural to begin by considering the one-dimensional analog of the two-dimensional problem. The basic problem then becomes: Given a sample $z_1, z_2, \cdots, z_n$ from an unknown, moderately normal, univariate distribution of known location, and given a function $\phi(z)$, which may be taken to be of the form e^{-cz^2}, how it is reasonable to estimate the average value of the function over the distribution? It is relevant, natural, and important to specify the distribution as moderately normal—relevant and natural since we would not have given much, if any, consideration to guidance from normal theory if the sample made it clear that the population was markedly non-normal; and important since only those procedures of estimation will be considered which would be acceptable if the population were in fact normal.

Note: References [18] through [27] report some of the approaches to the

problem of compound distributions, while Refs. [28] through [35] relate to compound normal distribution. The literature on contagious (and negative binomial) distributions has been omitted; many references can be traced from the lists given on pages 67–68, 105–6, 194, and 248–49 of Volume 14 of *Biometrics*.

4. Location and Scaling: Standard and Alternative Pencils

A pure location problem is one in which the specified cumulatives (cumulative distributions) are transformations carrying any distribution of the specification into any other. That is, the cumulatives are $G(z-\mu)$ where $G(z)$ is the prototype cumulative, and μ is a location parameter. A pure scaling problem is one in which the specified cumulatives are $G(z/\lambda)$, where $G(z)$ is again the prototype cumulative, and λ is a scaling parameter. The transformation $z = e^w, \lambda = e^\tau$ reduces any pure scaling problem to a pure location problem. Both cases provide estimation problems of the simplest form involving

1. A random sample from a one-dimensional distribution,
2. A one-dimensional parameter,
3. A group of transformations carrying each distribution into all the distributions concerned.

If a general situation can be persuaded to show itself in such an estimation problem, one may expect its essential nature to be most clearly revealed there.

We face a (location or) scaling parameter, changes in whose values sweep us back and forth along a one-dimensional family of distributions, in loosely geometric terms, along a *pencil* of distributions. In our general problem, which has been reduced and analogized to certain pure scaling problems, we must consider not one pencil, but many. We face a *nearly* normal distribution, but we do not know its precise shape in detail. Yet, with location fixed, each shape corresponds to a unique pencil, different pencils corresponding to different shapes. One of these pencils, that consisting of normal distributions, plays a unique and leading role. All methods of estimation to be considered must produce satisfactory results for the normal pencil.

But in reality we face not the normal pencil but an alternative pencil. The behavior of each method of estimation which is seriously considered must be studied on each of many pencils. This may be done one pencil at a time, but it will eventually be necessary to combine results across pencils.

The basic problem is to study the behavior *on the alternative pencil* of methods of estimation which are *correctly calibrated on the standard reference pencil* (in our instances the normal one).

5. Choice of Alternatives

Our aim is to study the behavior of estimators of location and scale, especially those optimized for normality, for a wider range of specifications. It would be possible to try to study their behavior in very varied circumstances, perhaps, even, for translations or scalings of an arbitrary distribution with a continuous density function. For the present this does not seem analytically manageable. Moreover, since gross non-normality will be detected, practical concern must be focussed on the effects of indetectable or barely detectable non-normality. We shall learn that such effects may be large.

THE FIRST ANSWER

The relative efficiency of the mean deviation will increase.

THE SECOND QUESTION

Will the relative efficiency ever reach 100%; in other words, will the mean deviation be as good a measure of scale as the standard deviation for any of the contaminated populations obtained by mixing two normal populations which have the same mean but whose standard deviations are in the ratio 3: 1?

NOTES: 1. The possible answers are "never reach," "just reach," "reach and go beyond."
2. The problem is a large-sample problem.
3. To find the answer, after giving the question long and careful thought, turn over two pages.

Thus it is natural to consider as simple a family of pencils as possible, selected to possess non-normality with large consequences and poor detectability. Convenience of computation is rather well served by confining attention to some family of compound normal distributions with two constituents, that is, to certain mixtures of two different normal distributions, whose two constituents can differ in scale, or in location, or in both.

Various special choices of family can be made. Different choices by different statisticians helped avoid agreement about the original problem. The writer, basing his choice partly upon empirical experience with largish "samples" of range-finder fluctuations and partly upon the great sensitivity of many procedures to "dirt in the tails," favored reference to "contaminated distributions" where the two constituents have equal means and different variances. R. D. Bennett tended to choose the case where the constituents had equal variances and unequal means. Since the qualitative nature of the results is opposite for these two cases, this choice is of considerable importance. Sections 7 and 15 will discuss this point further.

For the present, the prototype distributions will be *contaminated distributions* of contamination τ, at scale ratio h, where a fraction, τ, of the wider normal distribution, whose scale is h times that of the narrower normal

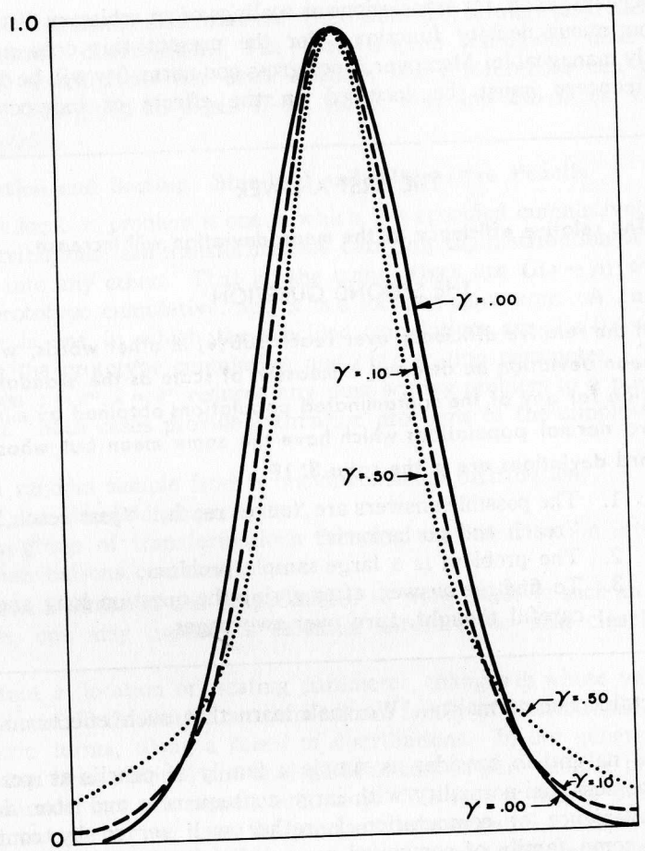

FIG. 1. Probability Densities of Normal and Contaminated Distributions (Central Densities Equated)

distribution, is combined with a fraction, $1-\gamma$, of the narrow one. The corresponding prototype cumulative is

$$N_{\gamma,h}(z) = (1-\gamma)\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,du + \gamma \int_{-\infty}^{z} \frac{1}{h\sqrt{2\pi}} e^{-u^2/2h^2}\,du\ ,$$

corresponding to a probability density

$$n_{\gamma,h}(z)\,dz = (1-\gamma)\frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz + \gamma \frac{1}{h\sqrt{2\pi}} e^{-z^2/2h^2}\,dz\ .$$

If $\gamma = 0$ or $\gamma = 1$, the corresponding pencil consists of normal distributions. Mainly because of experience with distributions of range-finder fluctuations, the value $h = 3$ was used throughout the investigations to be summarized

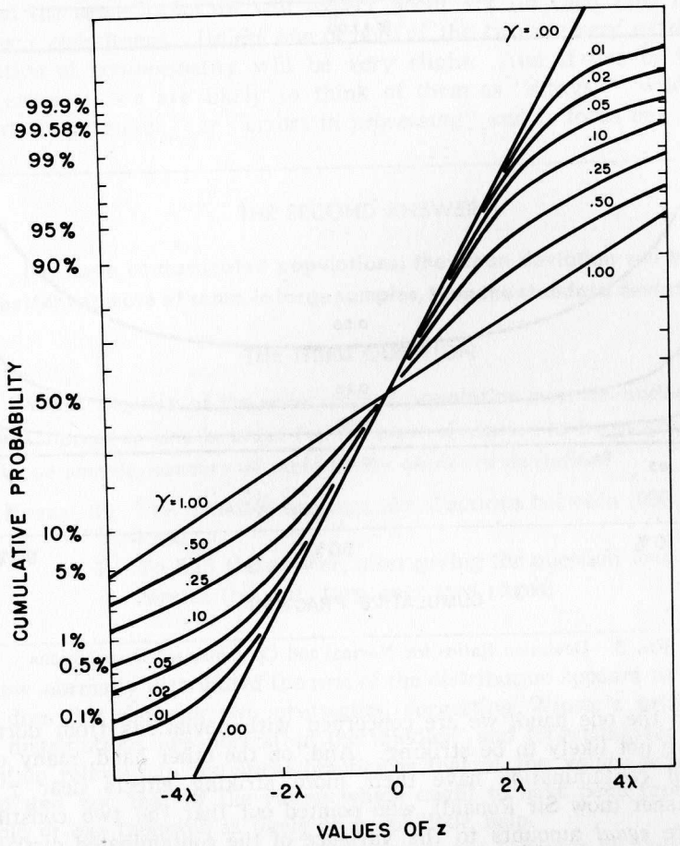

FIG. 2. Cumulatives of Normal and Contaminated Distributions

here. Thus the contaminated pencil will take the form $N_{\gamma,3}(z-\mu)$ in location problems and $N_{\mu,3}(ze^{-\tau})$ in scaling problems.

6. The Natural History of Contaminated Distributions

Before going on to consider the behavior of estimates based on samples from such contaminated distributions, it is well to learn something of the nature of the distributions themselves. This is probably best done graphically. Figures 1 and 2 show probability density functions (on a linear scale) and cumulatives (on probability paper), respectively, for a selected set of values of γ. It is quite clear from Fig. 1 that simple frequency distributions, even based on extremely large samples, are very unlikely to reveal the difference between $\gamma = .00$ and $\gamma = .10$ to the eye.

Throughout the discussion, just as in these figures, there will be continued emphasis on values of γ between .00 and .10. There are two reasons for

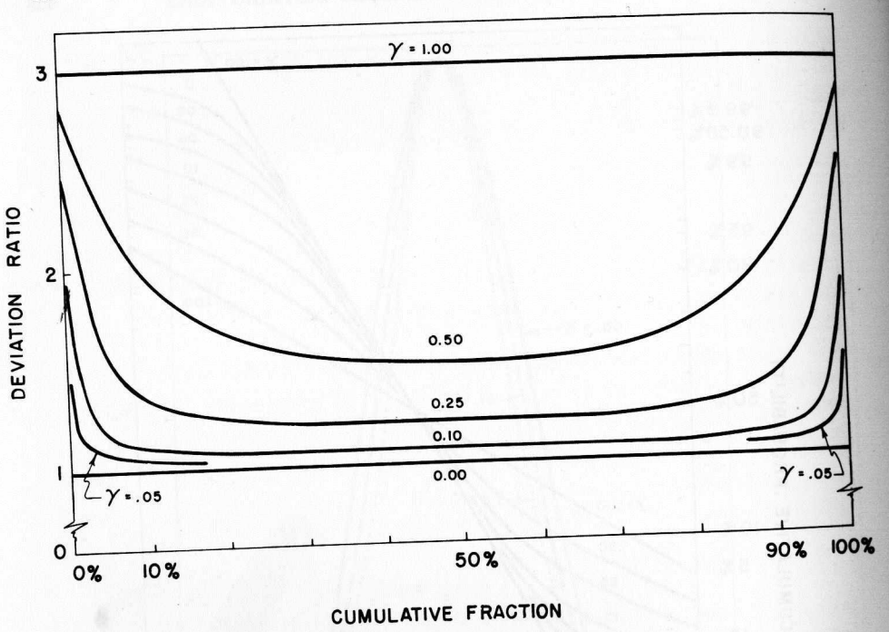

FIG. 3. Deviation Ratios for Normal and Contaminated Distributions

this. On the one hand, we are concerned with deviations from normality which are not likely to be striking. And, on the other hand, many of the effects of contamination have their more striking effects near $\gamma = .10$. R. A. Fisher (now Sir Ronald), who pointed out that the two constituents contribute *equal* amounts to the variance of the contaminated distribution when $\gamma = .10$, was the only statistician queried by the author to anticipate the relatively large effects of small γ. (In those figures where γ's larger than .10 appear, the scale of γ is linear in the proportions of the total variance contributed by the two constituents.)

Given a sample of several thousand from some distribution, how is it reasonable to inquire what value of γ might provide a contaminated distribution with "similar" behavior so far as tails are concerned? If such sample sizes seem unrealistic, consider a sample of 1000 from a population with $\gamma = .01$, which we shall learn is already large enough to produce behavior quite different from that for $\gamma = .00$ (i.e., that for normality). The average number of observations which come from the broader, rarer constituent will be only ten. Reference to Fig. 2 shows that the cumulatives do not deviate from one another by relevant amounts between -2.5λ and $+2.5\lambda$. Only some 40 percent of the observations from the wider component will fall outside these bounds, and we may thus expect only two observations from the upper tail of the broader, rarer constituent, and another two from the lower tail.

Beyond the same limits we will expect about six (in each tail) from the narrower constituent. Unless one or both of the two are very extreme, the indication of non-normality will be very slight. And if one or both are very extreme, we are likely to think of them as "strays," "wild shots," "errors in technique," or "errors in processing" and to focus our attention

THE SECOND ANSWER

For some contaminated populations, the mean deviation will be a better estimate of scale, in large samples, than the standard deviation.

THE THIRD QUESTION

What fraction of the wider normal population must be added to the narrower one in order for the mean deviation to be as good a large-sample measure of scale as the standard deviation?

NOTES: 1. The possible answers are fractions between .000 and 1.000.

2. To find the answer, after giving the question long and careful thought, turn over two pages.

on how normally distributed the rest of the distribution appears to be. (One who does this commits two oversights, forgetting Winsor's principle that "all distributions are normal in the middle," and forgetting that the distribution relevant to statistical practice is that of the values actually provided and not that of the values which ought to have been provided.) A sample of *one* thousand is likely to be of little help.

A convenient and effective way to display the behavior of the sample of several thousand (or of a theoretical distribution) is to plot the deviation ratio (fractile = % point):

$$\frac{\text{(deviation of fractile from mean)}}{\text{(deviation of same fractile from mean for normality)}}$$

against the cumulative fraction. For normality, the plot is a horizontal line. For contaminated distributions, the ends are raised. Figure 3 shows the results for several values of γ. Given a sufficiently large sample, this type of plot is quite revealing.

Finally, how can we conceive of contaminated distributions, or close approximation to contaminated distributions, arising in practice? Surveyors have long carefully distinguished between "errors," which afflict all their measurement, and "blunders," which affect only a small fraction and are usually so large as to be findable by checking procedures. Although surveyors' blunders are (i) not normally distributed, and (ii) more than three

times as large as surveyors' errors, this example suggests a model in which some source of fluctuation affects only a fraction, γ, of the observations. If this intermittent source supplies eight times the total variance supplied by all other sources, and all sources provide normally distributed contributions, the resulting distribution will be contaminated by fraction γ at scale ratio 3.

Other more general situations, where the scale of variability is heterogeneous, but where variation at each separate scale is normal, will produce distributions too long in the extreme tails for normality, distributions which clearly resemble contaminated distributions.

7. Long Tails or Short?

It is easy to think of many situations which produce distributions with longer tails than a normal distribution. (Heterogeneity of variance and causes acting only a small part of the time are, of course, the main possibilities.) But it is much more difficult to think of realistic situations which lead to tails shorter than a normal distribution. Sets of observations which have been de-tailed by over-vigorous use of a rule for rejecting outliers are inappropriate, since they are not samples. Quantitative characteristics of manufactured articles that have been sorted by gaging or measurement will tend to have short-tailed distributions, but, when the sorting is at fixed cut-offs, the mean of the sorted populations is not a location parameter, and their standard deviation is not a scale parameter.

A contribution taking two values, each with substantial probability, can shorten tails somewhat, as when it creates a mixture of two normal distributions with the same variance and different means. But, to be effective in tail-shortening, it must meet many requirements. It must be a large contribution, and it must be almost alone of its kind, for the sum of many such will converge to normality. Its values can only rarely be widely different from the two particular values, since intermediate values with substantial probability will reduce its effect greatly. (And a few unusually extreme contributions will erase its effect entirely.) The most tail-shortening it can contribute corresponds to $\gamma_2 = -2$ or $\beta_2 = 1$ (Fisher's and Pearson's measures of normality, respectively), while tail-lengthening contributions can provide much larger positive γ_2's.

It is not suprising that, in practice, long tails seem more frequent than short. As Box and Andersen say ([17], p. 2), "Published data are comparatively meagre but the frequency distributions given in the older issues of *Biometrika* give little ground for supposing that distributions usually follow the normal law." They go on to cite certain examples where large samples reveal the long-tailedness of populations. To their list should certainly be added Student's paper on the errors of routine analyses ([36], also in [37]), which provides much evidence, based on very many small samples, on the long-tailedness of experimental measurement. All in all, the experimental evidence seems to lead us to beware long tails as strongly as does our theoretical insight.

8. Asymptotic Techniques

By definition, large-sample results are asymptotic results. Asymptotic results have always tended to be either simple and heuristic or complex and rigorous.

The first step in attacking the problem of sampling from contaminated distributions was to set up a system of asymptotic techniques whose use could be both simple and rigorous [38], [39], [40]. Definitions and results are set forth in Appendix A. Four points deserve remark here:

(i) Both the asymptotic average value of $y\{Z_n\} = \text{ave}\vec{}\, g(Z_n)$ and the asymptotic variance of $g(Z_n) = \text{var}\vec{}\, y(Z_n)$ will exist for reasonable $g(z)$ whenever the distributions of the Z_n's converge in a reasonable way to a single point, as is almost always the case when the Z_n's are reasonable bases of estimate.

(ii) $\text{Ave}\vec{}\, Z_n$ and $\text{var}\vec{}\, Z_n$ are deliberately not defined exactly, in the sense that terms which go to zero faster than $\text{var}\vec{}\, Z_n$ itself may be freely added and subtracted.

(iii) Only the distributions of the individual Z_n appear in the definitions.

(iv) The so-called "Δ-process," or "method of propagation of error," is usually valid in terms of these asymptotic techniques.

In this connection it is interesting to note that Harold Hotelling once wrote ([41], p. 36): "Another example of nonrigorous mathematics used extensively in statistics is the whole business of asymptotic standard errors found by the differential method. It is desirable that good mathematics replace bad in such connections."

In the sections that follow, the word "asymptotic" will have the meaning that is defined in Appendix A.

9. The Location Problem

Since contaminated distributions studied here are symmetric, most estimates of location are unbiased for all amounts of contamination. Thus only variability need be considered, and the asymptotic variances of the various estimates will be used to indicate preferable choices.

If both the amount of contamination, γ, and the scale, λ, of the underlying distribution were known, one could be both simple and "efficient" (asymptotically) by using maximum likelihood ideas, either directly or in terms of Fisher's score (the logarithmic likelihood derivative). It is natural to ask how well the score appropriate for one fraction of contamination, γ', (and the correct value of λ) will do if it is used on observations drawn from a distribution with another fraction, γ, of contamination. Calculation shows ([6], Table 6) that, if we assume $\gamma' = .00$, we may suffer rather low efficiency (90% for $\gamma = .02^+$, 80% for $\gamma = .05$, 70% for $\gamma = .10$), but if we assume γ' between .01 and .10, we shall obtain at least 96% efficiency over the whole range $.00 \leqq \gamma \leqq .10$. Thus there is more to gain than to lose by assuming a moderate degree of contamination.

This result is not yet a practical answer for at least two reasons. The

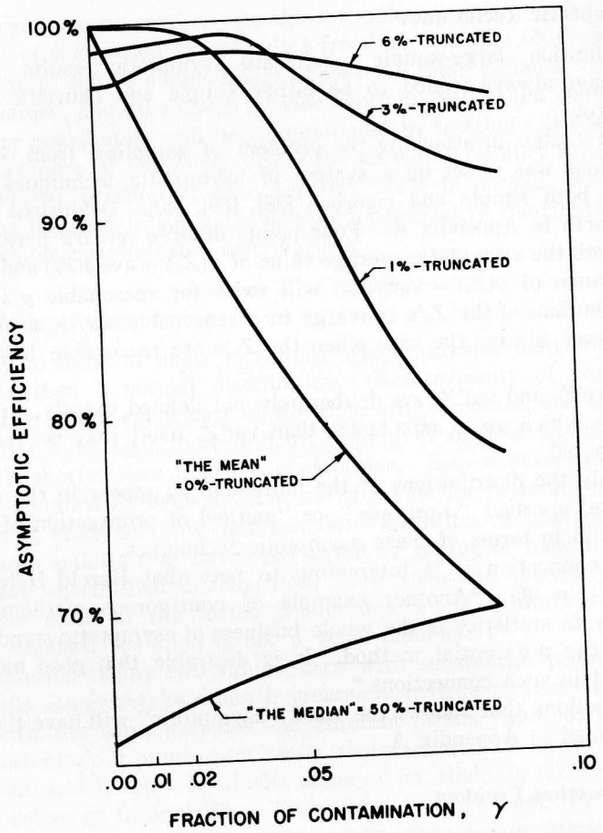

FIG. 4. Asymptotic Efficiency, for Location, of Truncated Mean

estimator studied was specifically constructed in terms of a specific contaminated distribution, so that we may rightfully distrust its efficiency for other types of deviation from non-normality. And it makes use of knowledge (except when $\gamma = .00$) of the value of λ which we would not have in practice.

Since our concern is with large samples and asymptotic results, the latter difficulty can be easily gotten over. Maximum likelihood technique can be replaced by the use of a weighted linear combination of order statistics, where the weights are chosen to approximate the derivative of the score, evaluated at the corresponding order-statistic expectations. Asymptotically, the corresponding estimating techniques will be equivalent (cf. [6] and Jung [42]).

When the resulting weighting functions were examined, it appeared that they could be rather well approximated by appropriate truncated means. A *p%-truncated mean* (sometimes called a *broadened median*) is the mean of the observations remaining when the $p\%$ highest and the $p\%$ lowest have

been set aside. The truncated means form a natural bridge between two classical estimates of location, since the 0%- and 50%-truncated means are the mean and the median, respectively.

THE THIRD ANSWER

When just less than .008 of the mixed population comes from the wider normal population, the mean deviation has the same large-sample precision as the measure of scale that the standard deviation has.

NOTE: Do not judge an answer which deviates widely from .008 harshly. Many distinguished and experienced statisticians have given answers between .150 and .250.

THE FOURTH AND FIFTH QUESTIONS

Can I expect to know whether or not the populations from which I draw large samples deviate from normality as much as, say, a contaminated population containing .008, .02, or .05 of the wider normal?

If I cannot, what should I do about estimating the scale of an actual population from a large sample?

NOTES: 1. To obtain insight into whether or not you may know, examine Fig. 1 and read Section 5.
2. To obtain insight into what to do, read Sections 11 to 13.

Figure 4 displays the asymptotic efficiencies of the 0%-, 1%-, 3%-, 6%-, and 50%-truncated means. Clearly, a small amount of truncation brings great reward for small loss. If 1% of the sample is deleted from each end, for example, the greatest loss is 0.4% and the gain may exceed 12%. If there is serious fear of contamination, the use of the 6%-truncated mean, with a maximum loss of 3% and a maximum gain of 26%, seems entirely reasonable.

Although the changes in efficiency here are small in comparison with those which we will face in considering scaling, they are not negligible, as they rise for substantial contamination to the level of the difference in efficiency, on normal theory, between the mean and the median. [See Jeffreys ([43], Sec. 5.7, pp. 287 ff.) for another approach to the location problem for non-normal samples.]

Investigation of the terms in $1/n^2$ indicates that, in samples as small as 100, the asymptotic formulas probably overstate variances by only a few percentage points [7].

10. Problems of Scaling Type: Bases of Scaling

In Section 4, λ was called *a* scaling parameter. Why the caution? Be-

cause once alternative pencils must be dealt with, there are many distinct numerical functions of distributions which can rightly be called scaling parameters. Consider, for example, the scaling of pencils consisting of distributions centered at the origin with finite second moments, and, in particular, consider the pencil of centered normal distributions and the pencil of centered logistic distributions.

The parameters α and β, defined by

$$\alpha^2 \doteq \text{ave } z^2 ,$$

$$\Pr\{z \geqq 3.090\,\beta\} = 0.001 ,$$

are both scaling parameters. Along any one pencil they enlarge and contract in strict proportion to the enlargement and contraction of the distribution. For the normal pencil their values agree. For the logistic pencil their values are in a constant ratio of 1.58. For any one pencil for which they are both finite, their values are in a constant ratio.

So long as any one pencil is considered, the estimation of α and the estimation of β are equivalent problems. Once two or more pencils are considered, the problems are no longer equivalent. For more than one pencil, there is no unique scaling problem; instead there are very many scaling problems.

The well-known fact that location seems to be simpler than scaling is reflected in the simplicity gained by adopting a logarithmic measure of scale. We shall write $\lambda = e^{\tau}$, and consider our contaminated pencils to be specified in the form $N_{3,\gamma}(ze^{-\tau})$. It is now simple to construct a variety of estimators of τ whose bias is independent of the actual value of τ. For instance:

(i) From the *pth degree absolute* (sample) *moment* $(\sum |z_i|^p)/n$ we can form

$$\frac{1}{p}\log\frac{1}{n}\sum |z_i^p| .$$

(ii) The *p%-interfractile distance* is the distance from the observation *p%* from the bottom of the sample to the observation *p%* from the top of the sample. Its logarithm has constant bias.

(iii) The *p%-truncated variance*, whose square root is the *p%-truncated standard deviation*, is the variance of those observations which remain when the *p%* highest and *p%* lowest have been set aside. The logarithm of the truncated standard deviation has constant bias.

In particular, we shall also wish to estimate the population average of e^{-cz^2}, which will be called a Gaussian average. The only *completely unbiased* (that is, unbiased for every possible distribution) estimate of such an average is the sample mean of the same expression. Hence, we consider the Gaussian means $(\sum e^{-cz_i^2})/n$ as a base of estimate. We can find a function of a given Gaussian mean which is asymptotically unbiased along the normal pencil. However, its asymptotic bias along a contaminated pencil will not be constant (although an asymptotically equivalent base of estimate does

have constant bias; see Section B.3 of Appendix B). Thus, even after changing to the logarithmic scaling parameter it may be sometimes necessary or desirable to deal with varying bias.

It will be helpful to notice (i) that it is often useful to write $e^{-b(z/\lambda)^2}$ for e^{-cz^2}, thus replacing c by the non-dimensional $b = c\lambda^2$, (ii) that as $b \to 0$, Gaussian means, considered as bases of estimates, behave more and more like the second moment (like the variance), (iii) that as $b \to \infty$, Gaussian means, considered as bases of estimate, behave more and more like *p%*-interfractile distances for $p\% \to 50\%$. (In the last instance, asymptotic variances approach ∞.)

11. Size of Sample

A statistical problem is usually called a large-sample problem when $1/n$ is so small that everything else can be neglected. If this is not the case, some will automatically call it a small-sample problem. Others will reserve the latter term for problems which (a) can be solved explicitly, or (b) are treated by experimental sampling. A more refined classification is needed in the present situation, where, by either definition, there is a small-sample problem. Many derived distributions can be written down in closed, but apparently impractically complicated, forms. Many specific problems can be effectively attacked for specific small-sample sizes by experimental sampling. To date, little if anything has been done in this direction. All the work reported here concerns large sample sizes.

But it will often be essential to recognize different "largenesses" of samples, since the relative importance of variance and bias will vary. Typically, asymptotic variability falls off with $1/n$, while asymptotic bias tends to a non-zero limit. For *very* large samples, then, bias outweighs variance, and the choice of estimator is primarily made to match bias, and only secondarily to reduce variance.

But for slightly large samples, it may well be that bias is unimportant in comparison with variability *and yet* asymptotic results are precise enough to be useful. A similar reliance on variability alone arises when the aim is to *compare* the scales of two or more distributions where, perhaps because of the similarity of their sources, it is reasonable to assume that all the distributions come from a single pencil (e.g., have the same value of γ). No matter how large the samples may be, they then act as if they were slightly large.

So long as estimators are considered whose bias (referred, say, to λ) is constant along a pencil, to consider "variability alone" means to consider variance. But, if bias varies along the pencil, it is necessary to make allowance for this fact, just as Mandel and Stiehler [44] have done in defining "sensitivity" and just as is exemplified by the more flexible forms of the Cramér-Rao inequality (e.g., [45], Sec. 32.3). The "effective variance" takes the form

$$\text{effective variance} = \frac{(\text{variance})}{\left[1 + \frac{d}{d\tau}(\text{bias})\right]^2} ,$$

so that the *asymptotic effective variance*, which is the appropriate measure, for this problem in our situation, becomes

$$\text{asymptotic effective variance} = \frac{\text{asymptotic variance}}{\left[1 + \frac{d}{d\tau}(\text{asymptotic bias})\right]^2}.$$

And, finally, for *moderately* large samples, it may be best to accept a carefully chosen amount of bias, in order to decrease variance enough to make a net reduction in some combined measure, such as the asymptotic average square deviation. (This latter case will be discussed only in Appendix B.)

12. Variability of Scaling

The bases of estimate considered in [8] were the following:

(i) pth-degree absolute moments for $0.2 \leqq p \leqq 5$,
(ii) Gaussian means for $0 \leqq c \leqq \infty$,
(iii) $p\%$-interfractile distances for $0.1\% \leqq p\% < 50\%$,
(iv) Selected truncated variances (2% and 5%).

Figure 5 shows the behavior of the product of asymptotic effective variance and sample size for $.00 \leqq \gamma \leqq .50$ and selected examples of each type. Clearly the second moment, which corresponds to the standard deviation, is the least safe of all. Its use can be recommended only when γ is far less than .01, and we are rarely sure of this. More detail is provided by Fig. 6, which shows, for the range $.00 \leqq \gamma \leqq .10$ and some of the better-behaved estimates, the percentage of excess of the effective variance of estimate over the Cramér-Rao lower bound ([45], Sec. 32.5). Probably the most promising of the alternatives shown, if substantial γ is to be feared, are (i) the mean of $e^{-z^2/4}$, and, if γ is quite likely to be $\leqq .07$, say, (ii) the 2%-truncated variance. In general, it would appear that a suitable truncated variance would do better over a narrow range of γ, while a suitable Gaussian mean would do better over a broader region. (It is almost certain that a procedure of varying truncation, which sets aside varying percentage of the observations from different samples, would do even better. Reasonable procedures do not seem easy to write down. The study of their asymptotic variability seems likely to be much more difficult.)

It is hard to imagine a situation where contamination would appear, and yet appear in such small amounts as to make the standard deviation either as good as the mean deviation or nearly as good as the 2%-truncated standard deviation, at least insofar as variability of scaling goes.

Because of practical questions of computing, we *may* find averaging e^{-cz^2} uncomfortable. If so, then the most reasonable solutions for this problem are:

1. The truncated variance (and its square root, the truncated standard deviation) with 2% to 5% of the observations deleted from each tail,
2. The mean deviation.

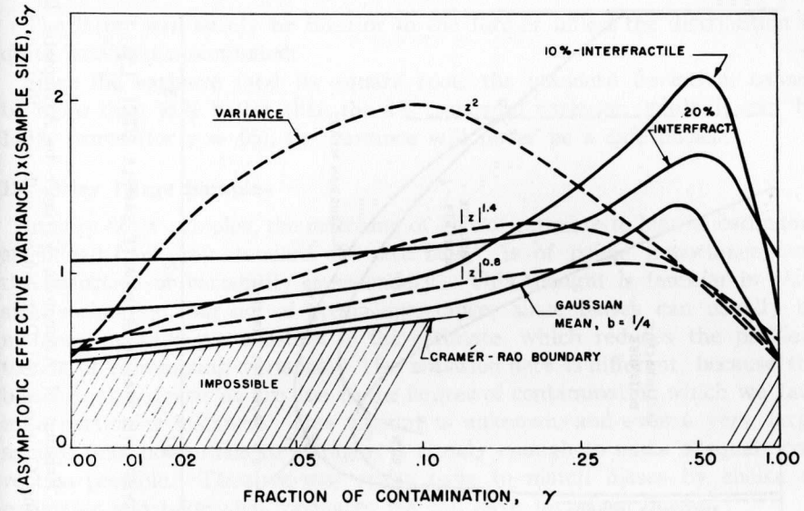

FIG. 5. Variability When Using Various Estimators in Scaling Slightly Large Samples

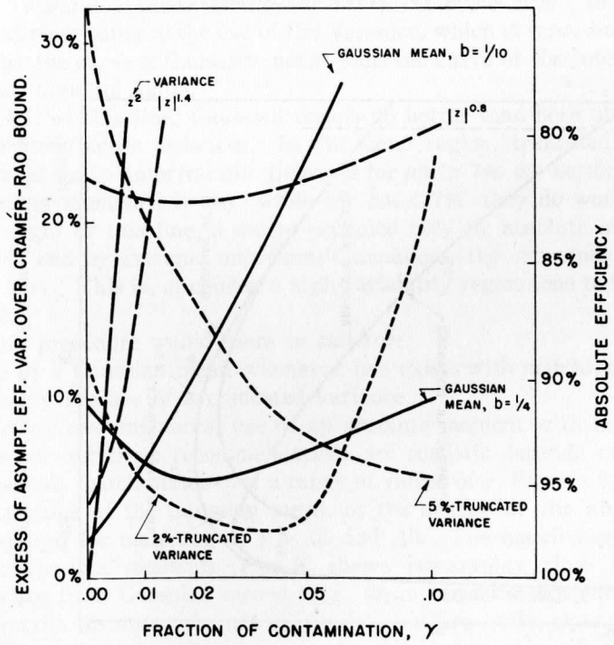

FIG. 6. Details of Variability When Using Various Estimators in Scaling Slightly Large Samples

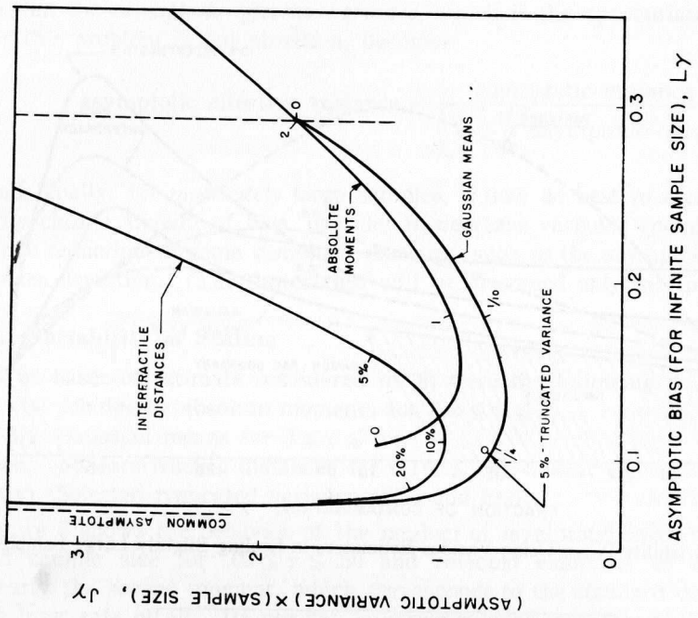

FIG. 8. Relation of Asymptotic Variance to Asymptotic Bias When Scaling Contaminated Distributions with $\gamma = .10$

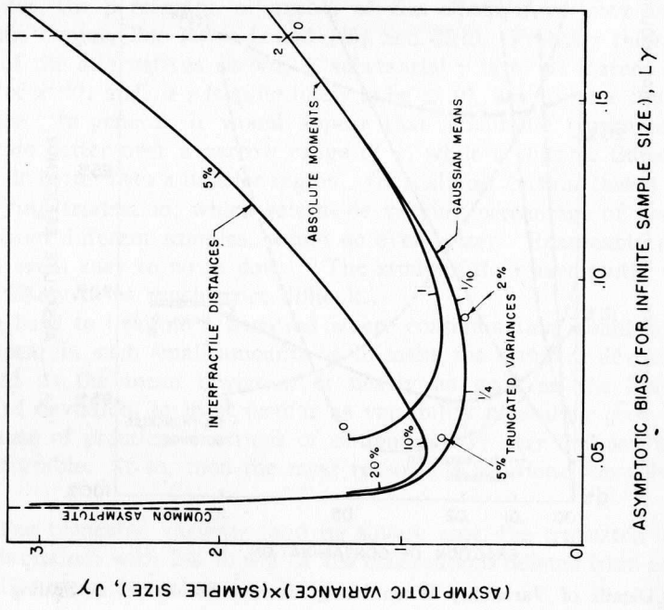

FIG. 7. Relation of Asymptotic Variance to Asymptotic Bias When Scaling Contaminated Distributions with $\gamma = .05$

The latter will surely be inferior to the former unless the distribution is quite heavily contaminated.

Since the variance (and its square root, the standard deviation) cannot be more than 11% better than the 2%-truncated variance, while it can be 140% worse (for $\gamma = .05$), the variance will never be a safe choice.

13. Very Large Samples

In very large samples, the matching of bias of estimate to bias of estimand, as judged from any standard of "zero bias," is of prime importance, and the reduction of variability is secondary. This thought is familiar in classical estimation, but not of great importance, since biases can usually be matched by adding a constant to the estimate, which reduces the problem to one of minimizing variance. The situation here is different, because the biases that concern us are due to the degree of contamination which we face in a particular instance. This amount is unknown, and even a very large sample may not suffice to estimate it closely enough to make adequate correction possible. Thus we may really have to match biases by choice of estimator and take what variances we can have for given biases.

We look, then, to plots of asymptotic variance, J_γ, against asymptotic bias, L_γ, to see what type of estimator gives the desired biases with least variance. Figures 7 and 8 do this for $\gamma = 5\%$ and $\gamma = 10\%$. In each case, the bias corresponding to the use of the variance, which is represented by the point where the curve of Gaussian means joins the curve of absolute moments, is a crucial dividing value.

To the left of this line, Gaussian means do better than both absolute-moment and interfractile distances. In this same region, truncated variances do almost as well. Interfractile distances for $p\% > 7\%$ do better than the corresponding moments, if any, while for $p\% < 7\%$ they do worse.

To the right of this line, a region occupied only by absolute moments of degree $\geqq 2$, and by extreme interfractile distances, the moments do better for fixed bias. This is, of course, a high-variability region, one to be avoided if possible.

Optimum procedure would seem to call for:

1. Use of a Gaussian mean whenever one exists with matching bias, or possibly the use of a truncated variance,
2. In other circumstances, use of an absolute moment with matching.

Whether or not these recommendations are realistic depends on whether or not one can match biases over a range of values of γ. Figures 9, 10, and 11 give the b-value of the Gaussian mean, or the degree of the absolute moment, required for matching at $\gamma = .05$ and .10. The match from Gaussian means to absolute moments (Fig. 9) shows remarkably close agreement. The matches from Gaussian means (Fig. 10) or absolute moments (Fig. 11) to interfractile distances are not so precise, but are quite close enough to make compromise quite effective.

14. The Reduced Form of the Original Problem

Location and scaling have been discussed in some generality, but what of

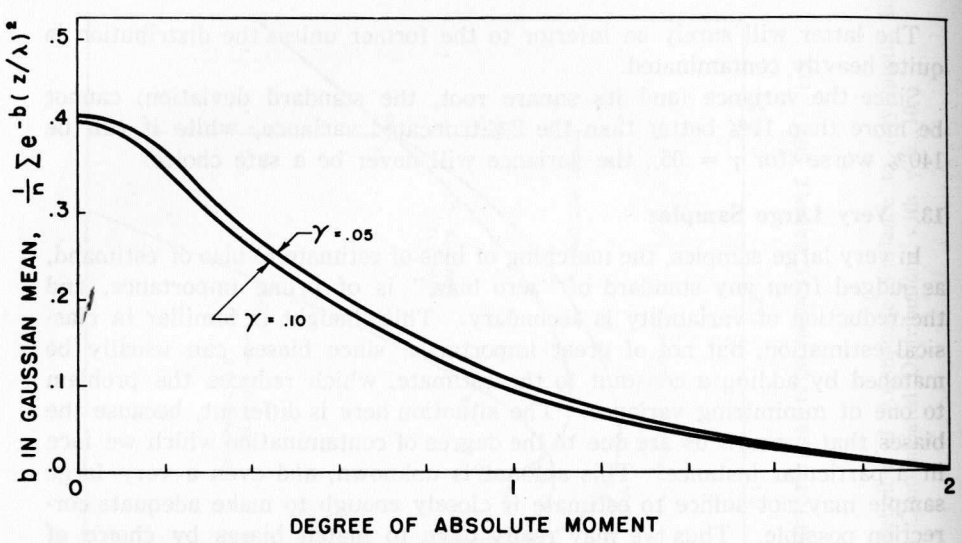

FIG. 9. Gaussian Means Matched in Bias to Moderate Absolute Moments

the special case which generated the study? Suppose that $b = 1$, so that the average value of $e^{-(z/\lambda)^2}$ is to be assessed from a sample which was drawn from a population known to be centered at zero.

The principal competitors to be considered are the mean deviation and the sample mean of $e^{-(z/\lambda)^2}$ itself. The resulting variances and biases are shown in Figs. 12 and 13. In slightly large samples, where variance dominates bias, the mean deviation is slightly preferred to the sample mean of $e^{-(z/\lambda)^2}$. Both are far better, if contamination is at all likely, than the sample variance (or standard deviation). In larger samples, where bias plays a controlling role, the sample mean of $e^{-(z/\lambda)^2}$ far outdoes its competitors.

The specific example treated was the average of the $e^{-b(z/\lambda)^2}$ for $b = 1$. In the original application, larger values of b would be more likely to occur than smaller values. If the same three competitors were considered for the estimation of such an average with $b > 1$, the relative asymptotic variances would be the same, and the relative asymptotic biases would increase in the same rank order as the worst asymptotic variances. Clearly, no one should dare to use the variance as a base of estimate, even though it is a sufficient statistic for precise normality.

15. How Far Have We Advanced?

How much better off are we after considering contaminated distributions than when we considered only the normal case? For a unique normal pencil, we have substituted a one-parameter family of contaminated pencils. This is still a very thin representation of all reasonable pencils, of all reasonable

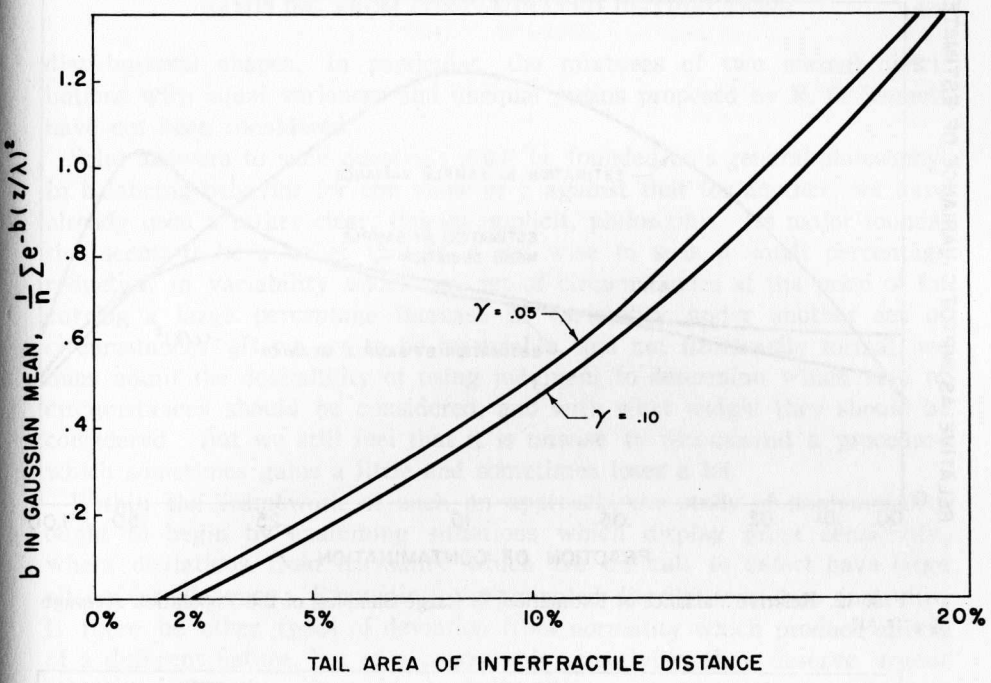

FIG. 10. Gaussian Means Matched in Bias to Moderate Interfractile Distances

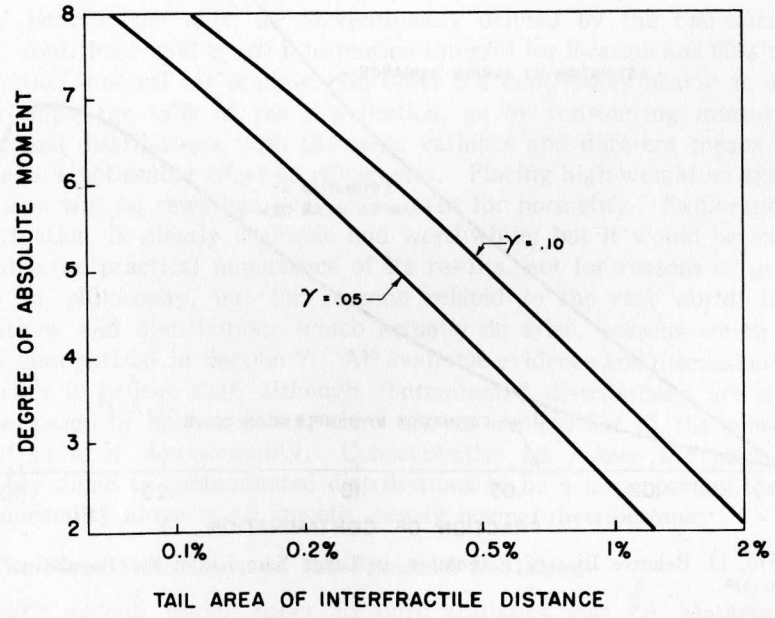

FIG. 11. Absolute Moments Matched in Bias to Extreme Interfractile Distances

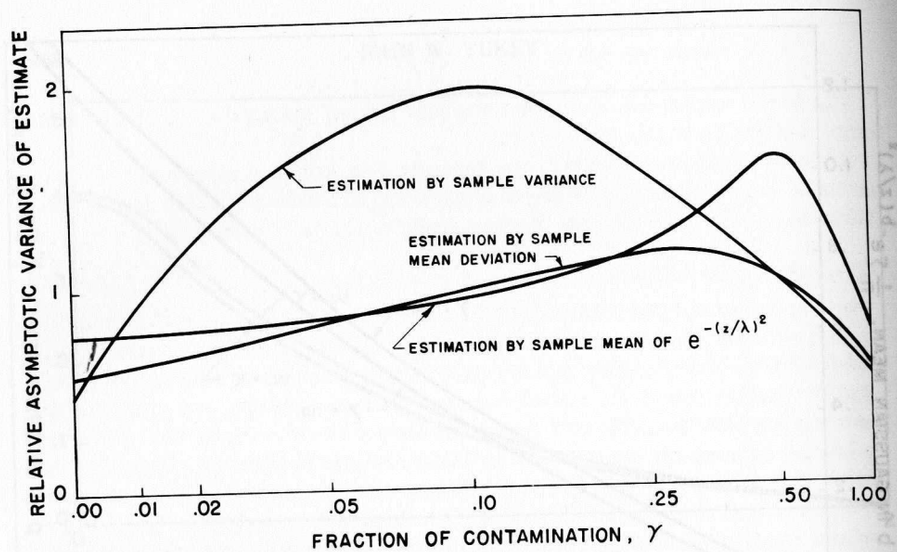

FIG. 12. Relative Variance of Estimation, in Large Samples, of the Population Average of $e^{-(z/\lambda)^2}$

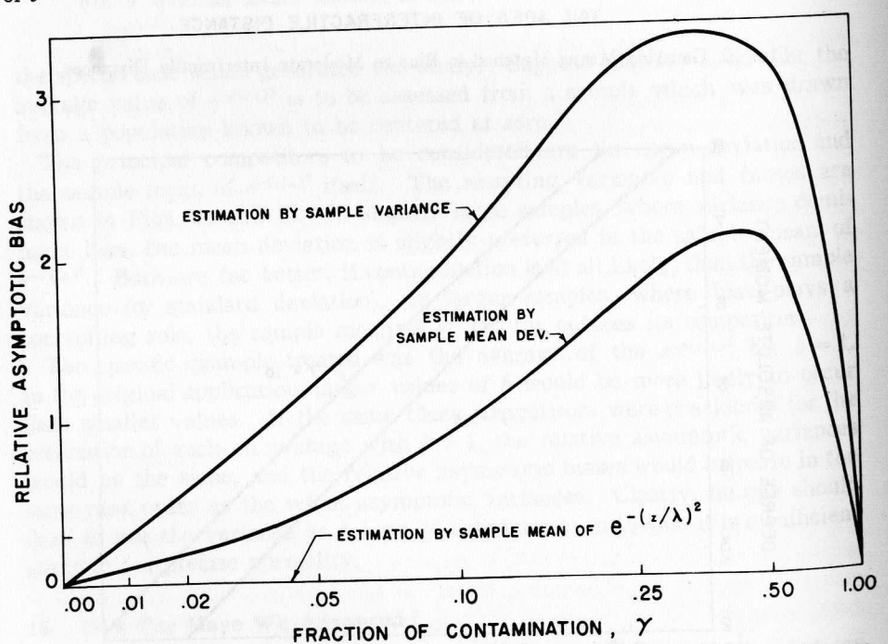

FIG. 13. Relative Bias of Estimation, in Large Samples, of the Population Average of $e^{-(z/\lambda)^2}$

distributional shapes. In particular, the mixtures of two normal distributions with equal variances and unequal means proposed by R. D. Bennett have not been considered.

Solid answers to such questions must be founded on a general philosophy. In balancing behavior for one value of γ against that for another, we have already used a rather clear, though implicit, philosophy. Its major foundation seems to be a belief that it is *not* wise to seek a small percentage reduction in variability under one set of circumstances at the price of incurring a large percentage increase in variability under another set of circumstances. If we are to be reasonable, and not necessarily formal, we must admit the desirability of using judgment to determine which sets of circumstances should be considered, and with what weight they should be considered. But we still feel that it is unwise to recommend a procedure which sometimes gains a little and sometimes loses a lot.

Within the framework of such an approach, the study of non-normality ought to begin by examining situations which display great sensitivity, where deviations from normality which are difficult to detect have large effects on efficiency. Examining contaminated distributions does just this. If there be other types of deviation from normality which produce effects of a different nature, but of a comparable sensitivity, they deserve urgent attention. But there is reason to doubt their existence, at least so long as we deal with smooth distributions.

Under slippage and expansion-contraction—in problems of location and scale—the tails of a distribution play a peculiarly important role. For normal pencils, the tails, as conventionally defined by the two-sided 5% points, contribute 28% to the information integral for location and 60% to the information integral for scaling. No other 5% contributes nearly as much.

Shortening the tails of the distribution, as by considering mixtures of two normal distributions with the same variance and different means, will also have a noticeable effect on efficiencies. Placing high weight on extreme deviations will be rewarded even more than for normality. Exploration of this situation is clearly desirable and worthwhile; but it would be easy to overvalue the practical importance of its results, not for reasons of general theory or philosophy, but for reasons related to the real world, to the populations and distributions which actually do arise, reasons which were briefly summarized in Section 7. All available evidence and discussion leads the writer to believe that, although contaminated distributions are a thin representation of non-normality, they are an exploration of the most important area of non-normality. Consequently, he judges the step from normality alone to contaminated distributions to be a large part of the step from normality alone to all smooth, nearly normal distributions.

16. The Mean Deviation and the Standard Deviation

Fisher's second major paper in pure statistics was "A Mathematical Examination of the Methods of Determining the Accuracy of an Observation

by the Mean Error and by the Mean Square Error" [46] (also in [47]). This paper, as Fisher says ([47] p. 2.757a), arose from an examination of a statement by Eddington. In a footnote to page 762 of Fisher's paper, Eddington acknowledges the need for Fisher's correction adding, however, "I think it accords with the general experiences of astronomers that, for the errors commonly occurring in practice, the mean error is a safer criterion of accuracy than the mean square error ...." The present investigation seems to support Eddington's conclusions, rather than Fisher's. We shall shortly need to ask why.

In his paper Fisher showed clearly, for the case of *exact* normality of distribution, that the standard deviation gave a much more precise estimate of scale than did the mean deviation. Moreover, he treated their joint distribution and pointed out ([46], p. 769) that the standard deviation was, in modern language, sufficient. (This is probably the first appearance of this concept.) He went on to point out that, if the distribution resembles a double exponential distribution rather than a normal distribution, it did not seem unreasonable to employ the mean deviation as an estimator of scale. The problem of bias did not evade him, since he pointed out that the factor connecting the average sample mean deviation with the population standard deviation was 12% greater for double exponential distributions than for normal distributions.

What did Fisher not do that we have done? Why did he reach a conclusion almost opposite to ours? First, he assumed that the purpose of scaling was the estimation of the population standard deviation, and this alone. (The assumption may well be far more reasonable in the astronomical applications he contemplated than in the situation which started the investigation surveyed here.) Second, he did not investigate in what sense, and to what degree, distributions would have to resemble the double exponential in order to make the use of the mean deviation preferable.[2] It is, I believe, fair to say that his analysis was clearly not carried far enough but that it contained the essential bases for a mathematical justification of Eddington's statement.

In his 1922 paper, "On the Mathematical Foundations of Theoretical Statistics," ([49], p. 314), Fisher said: "As regards problems of specification, these are entirely a matter for the practical statistician, for those cases where the qualitative nature of the hypothetical population is known do not involve any problems of this type. In other cases we may know by experience what forms are likely to be suitable, and the adequacy of our choice may be tested *a posteriori*. We must confine ourselves to those forms which

[2] He did state that β_2 seemed well fitted to provide an indication of this, remarking that $\beta_2 = 3$ for the normal, but $\beta_2 = 6$ for the double exponential. For contaminated distributions with $\gamma = .02$ at scale ratio 3, the value of β_2 is 5.75. Thus Fisher's criterion would suggest using the mean deviation for γ greater than about .022, a value we realize that we would be unable to discriminate from $\gamma = .00$ with the aid of a random sample of one thousand observations. See Burrau [48] for the use of the mean deviation.

we know how to handle, or for which any tables which may have been necessary have been constructed. More or less elaborate forms will be suitable according to the volume of the data. Evidently these are considerations the nature of which may change greatly during the work of a single generation ...." How should these words be translated, now that more than a generation has passed? The *a posteriori* testing of the "goodness of our choice" is now seen to require much more than a simple test of goodness of fit. There is a responsibility to explore, empirically or mathematically, all specifications which would be found to fit adequately, if tested, and to determine to what extent our conclusions, both about the situation underlying the data *and* about the statistical techniques most likely to illuminate such a situation, depend on which of the adequately fitted specifications we select. We are not yet able to shoulder the whole of this responsibility, but we badly need to acknowledge its existence.

In 1922, the theoretical specification of a population was almost always made in terms of a Pearson curve. Today, the Pearson curves serve us frequently as useful approximations to theoretical sampling distributions, and only rarely (except for the normal curve) as specifications. A wide variety of non-parametric techniques which use very broadly formulated specifications have been invented, investigated, and exploited. The time has come to study problems where the specification is neither so precise as a Pearson curve nor so diffuse as all distributions with a continuous density function. The precise formulation of such specifications as "the underlying distribution is nearly normal" in such a way as to be precise, manageable, and relevant to practice is now both important and overdue. It is time to return to the problem of specification.

17. Conclusions

What then are reasonable conclusions to draw from these investigations? Certainly these:

1. Problems of robustness of efficiency are probably as important as problems of robustness of validity, and, because of their relatively undeveloped state, deserve even more attention from statisticians.

2. Unless there is good reason to believe in its robustness, there is little if any sense in paying attention to an efficiency figure to greater precision than $\pm 10\%$ of itself.

3. In assessing the relative merits of competitive estimators in very large, or moderately large, samples, the ultimate purpose of the estimate is important, because biases due to apparently small mis-specifications of the problem (in our instances, changes in shape of distribution) cannot be neglected.

4. In large samples the sample mean is not nearly so safe an indicator of location as is the mean of the observations which remain after a small percentage of the highest, and an equal percentage of the lowest, have been set aside (use of a lightly truncated mean).

5. In slightly large samples, there is ground for doubt that the use of the variance (or the standard deviation) as a basis for estimates of scaling type is ever truly safe.

6. In moderately or very large samples, where matching of bias is important, the variance or standard deviation is safely used only to estimate one of the following (among all the estimands we have considered):

a. The variability (i) of a component, (ii) from a source, or (iii) of a value entering, as a datum, into a least-squares analysis,

b. Fractiles (% points) for tail areas between 1.5% and 2.5%.

Note: The tables of cell values entering an analysis of variance are rarely in any sense a large sample. When they are, the robustness of the associated significance and confidence procedures (cf. Pitman [15], Welch [16], and later developments) provides a good excuse for the use of means and variances in such analyses of cell values. If individual cell values themselves summarize large samples, then means or variances of these samples may well be replaced, as such summaries, by appropriately modified estimators.

7. Nearly imperceptible non-normalities may make conventional relative efficiencies of estimates of scale and location entirely useless.

8. If contamination is a real possibility (and when is it not?), neither mean nor variance is likely to be a wisely chosen basis for making estimates from a large sample.

9. As an interim measure, the use of truncated variances is likely to be quite satisfactory.

10. In smaller samples, the use of the mean deviation may be a frequently useful compromise,

11. More work is needed on the specific problem considered here, especially (i) in smaller samples, (ii) on lower bounds for J_γ when L_γ is given, and (iii) on a wider variety of estimators, including weighted order-statistic analogs of Gaussian means, and schemes of variable truncation.

Appendix A: ASYMPTOTIC TECHNIQUES

A.1 Possibilities

What can be expected from an organized asymptotic technique? At most, that it provides us with all terms through a given order, say $1/n$. The typical result might be of the form

$$\text{criterion} = A + \frac{B}{n} + o\left(\frac{1}{n}\right)$$

where $o(1/n)$ symbolizes additional terms of which we know only that they tend to zero faster than $1/n$. Or it might be of a more complex form, with an error term $o(1/n^2)$, and so on. But, if the algebraic manipulations are to be reasonable, it is almost certain that we must be left facing an error term of which we know only that it vanishes faster than some given power of $1/n$. This is the necessary nature of a simple asymptotic result.

If the technique is to be generally applicable, reflection shows that it cannot draw a sharp line between terms of order precisely $1/n$ and terms of higher order. Since, in particular,

$$7 + \frac{13}{n} + 97e^{-n}, \qquad 7 + \frac{13}{n} - \frac{1400}{n^2}, \qquad \text{and} \qquad 7 + \frac{13}{n}$$

all differ by $o(1/n)$, an asymptotic technique which goes only to $o(1/n)$ cannot appropriately distinguish among these three. It is thus convenient to take as asymptotic values not single numerical sequences but whole equivalence-classes of sequences.

The results so far described have been obtained and stated in terms of asymptotic average values and variances which are only *defined* up to $o(1/n)$. Consequently, their errors for any fixed value of n are unknown. (This need not prevent us from substituting finite values of n to obtain numerical illustrations.) But these errors will surely become negligible in comparison with $1/n$ for sufficiently large n, whatever definition and whatever evaluation of the expressions we take.

A.2 Adequate Intervals and Asymptotic Moments

Consider any sequence of chance quantities $\{Z_n\}$ and any finite open interval, I. Consider the *restrictions* $\{Z_n \,||\, I\}$ of the Z_n to I. $Z_n \,||\, I$ has as its distribution that part of the distribution of Z_n which is on I, inflated by the factor required to make the total probability of $Z_n \,||\, I$ in I equal to unity.

The bounded open interval I is *adequate* for $\{Z_n\}$ if the key condition,

$$\Pr\{Z_n \text{ not in } I\} = o(\operatorname{var}\{Z_n \,||\, I\}),$$

holds. If I' and I'' are both adequate for $\{Z_n\}$, (i) the ratio of $\operatorname{var}\{Z_n \,||\, I\}$ to $\operatorname{var}\{Z_n \,||\, I'\}$ tends to unity, and (ii) the common part of I' and I'' is also adequate for $\{Z_n\}$.

If s_n^2 is *any* sequence of positive numbers such that

$$\frac{s_n^2}{\operatorname{var}\{Z_n \,||\, I\}} \to 1$$

for any one adequate I, and hence for all adequate I's, the *asymptotic variance* of $\{Z_n\}$, written $\overrightarrow{\operatorname{var}} Z_n$, is defined to be s_n^2. (More precisely, the asymptotic variance is the equivalence class of such sequences, which consists of all t_n^2 with $t_n^2/s_n^2 \to 1$, but the looser language will be simpler, and safe enough to use.)

A notion of *asymptotic average value* can be defined by analogy with this definition. Its values are again equivalence classes of sequences, although again we shall speak of them as if they were sequences. We shall write $\overrightarrow{\operatorname{ave}}\{g(Z_n)\} = g_n$, if there is an adequate interval I on which $g(z)$ is bounded, and for which $\overrightarrow{\operatorname{ave}}\{g(Z_n \,||\, I)\} = g_n + o(s_n^2)$, where $o(s_n^2)$ means a quantity which

tends to zero faster than s_n^2. If this relation holds for one such I, then it holds for all such I [i.e., all those (i) which are adequate for $\{Z_n\}$, and (ii) on which $g(z)$ is bounded].

Given a sequence $\{Z_n\}$ of chance quantities, what can be said about the structure of the family of bounded open intervals which are adequate for $\{Z_n\}$? Because the common part of two adequate I's is itself adequate, there are only eight possibilities, which fall in three groups:

1. There are no adequate I's.
2. There exist numbers s and t, with $s < t$, such that an I is adequate if, and only if, it includes (a) the closed interval $[s, t]$, (b) the half-open interval $[s, t)$, (c) the half-open interval $(s, t]$, or (d) the open interval (s, t).
3. There exists a number z_0 such that an I is adequate if, and only if, it (a) includes z_0, (b) includes z_0 or has z_0 as left-hand end point, or (c) includes z_0 or has z_0 as right-hand end point.

In case (1) our machinery is not applicable at all. In cases (2) the sequence $\{Z_n\}$ does not behave the way a sequence of estimates or (normalized) bases of estimate should behave in large samples, though some of our machinery is applicable. The cases that really concern us are cases (3), where we say that $\{Z_n\}$ has asymptotic variance s_n^2 *near* z_0, written $\overrightarrow{\text{var}}\{Z_n\} = s_n^2$ near z_0.

There is a further simple characterization of cases (3). These are the cases where there are arbitrarily short adequate intervals. Thus "$\{Z_n\}$ has some asymptotic variance near some point" and "there exist arbitrarily short intervals which are adequate for $\{Z_n\}$" are equivalent statements, either of which characterizes a kind of behavior which is, in particular, shown (in large samples) by respectable estimates.

If $\{Z_n\}$ has asymptotic variance near z_0, then $\overrightarrow{\text{ave}}\{Z_n\} = m_n$ always exists, and converges to z_0. Clearly we should call m_n [again defined up to $o(s_n^2)$] the asymptotic average value of $\{Z_n\}$. It is natural to ask about higher asymptotic moments. The curious result is that the asymptotic moments of $\{Z_n\}$ about $\{m_n\}$ of order greater than two all "vanish" [i.e., are all $o(s_n^2)$, all belong to the zero equivalence class],

$$\overrightarrow{\text{ave}}\{(Z_n - m_n)^p\} = 0\,, \qquad p > 2\,,$$

and, as a corollary

$$\overrightarrow{\text{ave}}\{(Z_n - z_0)^p\} = (m_n - z_0)^p\,, \qquad p \neq 2\,.$$

It might seem that, if asymptotic moments (about the asymptotic mean) of order higher than two vanish, leaving only the asymptotic mean and the asymptotic variance, at least in the present sense of "asymptotically," all trace of shape of distribution is lost. This proves not to be the case.

A.3 Shapes, and Their Justification

A distribution is standardized if it has zero mean and unit variance. The distribution of $(Z - \text{ave}\{Z\})/(\text{var}\{Z\})^{\frac{1}{2}}$ is standardized, provided $\text{ave}\{Z_n\}$ and $\text{var}\{Z_n\}$ exist. Thus, if $\Pr\{Z_n \text{ in } I\}$ is never zero, the distributions of the chance quantities

$$\frac{(Z_n \,||\, I) - \text{ave}\{(Z_n \,||\, I)\}}{(\text{var}\{Z_n \,||\, I\})^{\frac{1}{2}}}$$

are all standardized. This sequence of standardized distributions may converge, or it may not; it may converge to a distribution, or it may not. If it does converge to a distribution, the limiting distribution is naturally regarded as expressing the *limiting shape* of $\{Z_n \,||\, I\}$. (Note that it is completely determined.)

If I and I' are both adequate for $\{Z_n\}$, and $\{Z_n \,||\, I\}$ has a limiting shape, then $\{Z_n \,||\, I'\}$ has the same limiting shape. Thus it is appropriate to call this common limit the asymptotic limiting shape of $\{Z_n\}$.

Limiting shapes, ordinary or asymptotic, need not be expressed by standardized distributions. Any distribution with $\text{ave}\{z^2\}$ *less than or equal to* 1 and $\text{ave}\{z\}$ *equal to* 0 can be a limit of standardized distributions. Only those with $\text{ave}\{z^2\}$ *equal to* 1 are standardized or, as we shall now say, represent *faithful shapes*. The others represent *shrunken*, or *degenerate shapes* (they have $\text{ave}\{z\} = 0$ and $\text{ave}\{z^2\} < 1$). Standardized distributions can converge either to a faithful shape or to a shrunken shape.

In the other direction, a sequence of standardized distributions cannot be wholly divergent, for any such sequence must contain a convergent subsequence. Thus, if a sequence of standardized distributions does not converge, it contains subsequences converging to two or more different limiting shapes. The limiting *shape behavior* of a sequence of distributions is thus determined by specifying (i) those limiting distributions toward which at least one (standardized) subsequence converges, and (ii) what subsequences converge to which of these limiting distributions. If all these limiting distributions represent faithful shapes, then the limiting behavior is *non-shrinking*. A special case of *non-shrinking limiting behavior* occurs when the standardized distributions converge to a single distribution of unit variance.

The result about the limiting shapes of $\{Z_n \,||\, I\}$ and $\{Z_n \,||\, I'\}$ can now be extended to assert that these two sequences have identical limiting behavior when I and I' are both adequate. The *asymptotic limiting shape behavior* of $\{Z_n\}$ can thus be meaningfully defined whenever some interval I is adequate for $\{Z_n\}$.

We can now state results which show to what extent the notions "asymptotic average value" and "asymptotic variance" are extensions of the notions "ordinary average value" and "ordinary variance." If $\mu_n = \text{ave}\{Z_n\} \to z_0$ and $\sigma_n^2 = \text{var}\{Z_n\} = 0$, *and if* the limiting shape behavior $\{Z_n\}$ is non-shrinking, then $\overrightarrow{\text{ave}}\{Z_n\} = \mu_n$ and $\overrightarrow{\text{var}}\{Z_n\} = \sigma_n^2$ near z_0. (Examples show that the hypothesis on the limiting behavior cannot be eliminated.) As a simple consequence, it follows that, if some I is adequate for $\{Z_n\}$ whose limiting behavior is non-shrinking, and for which $\text{var}\{Z_n \,||\, I\} \to 0$ and $\text{ave}\{Z_n \,||\, I\} \to z_0$, then $\{Z_n\}$ has asymptotic variance $\overrightarrow{\text{var}}\{Z_n \,||\, I\}$ near z_0. In particular, if $\text{var}\{Z_n\}$ exists and converges to zero, then $\overrightarrow{\text{var}}\{Z_n\} = \text{var}\{Z_n\}$.

A.4 Propagation Results

The next group of results relate the behavior of $\{g(Z_n)\}$ to that of $\{Z_n\}$. It is useful to do this under as weak hypotheses on $g(z)$ as possible. Thus parallel results with differing hypotheses may be desirable.

The *typical conclusion* is that

$$\text{ave}\,\{g(Z_n)\,||\,J\} = g(z_n) + o(s_n)\,, \qquad \text{var}\,\{g(Z_n)\,||\,J\} = A^2 s_n^2 + o(s_n^2)\,,$$

where $A = g'(z_0)$, and where, moreover, if $A \neq 0$, $\vec{\text{var}}\,\{g(Z_n)\} = A^2 s_n^2$ near z_0. This conclusion follows from the hypotheses:

(i) $g'(z)$ exists and is continuous (and not infinite) in an interval I continuing z_0, and

(ii) $\vec{\text{var}}\,\{Z_n\} = s_n^2$ near z_0,

or from the hypotheses, weaker for $g(z)$, but stronger for $\{Z_n\}$;

(i′) $g'(z_0)$ exists and is not infinite (only at the point z_0!),

(ii′) $\vec{\text{var}}\,\{Z_n\} = s_n^2$ near z_0, and

(iii) $m_n = \text{ave}\,\{Z_n\,||\,I\} = z_0 + o(s_n)$.

These latter hypotheses are also sufficient to establish that the asymptotic limiting shape behavior of $\{g(Z_n)\}$ is the same as that of $\{Z_n\}$.

The latter two results can be put in the following form: Provided that

(i) $A = g'(z_0)$ exists finite and unequal to 0,

then

(ii) $\vec{\text{var}}\,\{Z_n\} = s_n^2$ near z_0,

(iii) $\vec{\text{ave}}\,\{Z_n\} = z_0 + o(s_n)$, and

(iv) the asymptotic limiting shape behavior of $\{Z_n\}$ being so-and-so;

together imply that

(ii_g) $\vec{\text{var}}\,\{g(Z_n)\} = A^2 s_n^2$ near $g(z_0)$,

(iii_g) $\vec{\text{ave}}\,\{g(Z_n)\} = z_0 + o(s_n)$, and

(iv_g) the asymptotic limiting shape behavior of $\{Z_n\}$ being so-and-so;

and, indeed, (ii) and (iii) together imply (ii_g) and (iii_g).

Thus the class of sequences of chance quantities which satisfy analogs of (ii) and (iii) is closed under those changes of $\{Z_n\}$ to $\{g(Z_n)\}$ for which $g'(z_0)$ exists finite and non-zero. And the same is true for the subclasses defined by particular asymptotic limiting behavior.

Generalizations of these results to the case of $\{g_n(Z_n)\}$ are given in [40], which also contains details and proofs of the material summarized here.

Appendix B: THE EXPRESSION AND BALANCING OF BIAS AND VARIANCE

B.1 The General Formulation

The general formulation will be described in notation that (i) applies immediately to the specific instances of sampling from contaminated distributions and (ii) can also be interpreted as applying to general instances of estimation in the alternative pencil. Asymptotic operators without subscripts, like $\vec{\text{ave}}$ and $\vec{\text{var}}$, refer to evaluations for the reference pencil, in

our special instances a pencil of normal distributions. Asymptotic operators with subscript γ, like $\vec{\text{ave}}_\gamma$ and $\vec{\text{var}}_\gamma$, refer to evaluations for the alternative pencil, in our special instances a pencil of distributions each contaminated by a fraction γ at scale ratio 3.

If $u = u_n = u(z_1, \cdots, z_n)$ is a function of observations suitable as a base of estimate for the parameter θ, then, for a suitable $d(\theta)$, $\vec{\text{ave}}\,u_n - d(\theta)$ and $\vec{\text{var}}\,u_n$ both tend to zero like $1/n$. It is easy to describe ([8], Sec. 9) how to form a function t of u such that

$$\vec{\text{ave}}\,t = \theta\,, \qquad \vec{\text{var}}\,t = \frac{1}{n} i(\theta)\,.$$

This new function of observations estimates θ in the reference pencil of distributions. In the alternative pencil, we have

$$\vec{\text{ave}}_\gamma\, t = H_\gamma + \frac{1}{n} K_\gamma\,, \qquad \vec{\text{var}}_\gamma\, t = \frac{1}{A} J_\gamma\,,$$

where all of H_γ, J_γ, K_γ may depend on θ, and where manageable general expressions can be given for these three functions of θ ([8], Sec. 10).

If bias is not a concern, then the appropriate measure of variability of t as an indicator of θ in the alternative pencil is related to the sensitivity of t as defined by Mandel and Stiehler [44]. The appropriate asymptotic measure of variability is, explicitly,

$$G_\gamma = \frac{J_\gamma}{(H_\gamma')^2}\,,$$

where $H_\gamma' = (d/d\theta) H_\gamma$.

If both bias and variability are involved, it is convenient to put $H_\gamma = \theta + L_\gamma$ so that

$$\vec{\text{ave}}\,t = \theta + L_\gamma + \frac{1}{n} K_\gamma\,, \qquad \vec{\text{var}}\,t = \frac{1}{n} J_\gamma\,.$$

Here L_γ indicates the asymptotically important part of the bias, and J_γ is now the more appropriate measure of variability, since, referred to θ,

$$\text{average square deviation} = \vec{\text{var}}_\gamma\, t + (\vec{\text{ave}}_\gamma\, t - \theta)^2 = \frac{1}{n} J_\gamma + L_\gamma^2 + \frac{2}{n} L_\gamma K_\gamma\,,$$

while if it is referred to $\theta + L_{e\gamma}$, where $L_{e\gamma}$ is the asymptotic bias of the quantity [e.g., the average of $\phi(z)$] which we wish to assess, then

$$\begin{aligned}\text{average square deviation} &= \vec{\text{var}}_\gamma\, t + \vec{\text{ave}}_\gamma\,(t - \theta - L_{e\gamma})^2 \\ &= (L_\gamma - L_{e\gamma})^2 + \frac{1}{n}[J_\gamma + 2(L_\gamma - L_{e\gamma}) K_\gamma]\,.\end{aligned}$$

Any serious attempt to reduce the average square error by choosing a basis of estimate from a one-parameter family of possibilities will have to make $L_\gamma - L_{e\gamma} \to 0$ as $n \to \infty$, so that $(1/n)(L_\gamma - L_{e\gamma}) K_\gamma$ will be $o(1/n)$, and

$$\text{average square deviation} = (L_\gamma - L_{e\gamma})^2 + \frac{1}{n} J_\gamma$$

since the left-hand side is defined only up to $o(1/n)$.

B.2 Balancing Bias and Variance

In order to discuss the balancing of bias and variance by the choice of a base of estimate out of a one-parameter family of bases of estimate, we introduce a somewhat peculiar assumption. It is that some rule has been given so that $\vec{\text{ave}}_\gamma t$ and $\vec{\text{var}}_\gamma t$ are determined to $o(1/n^2)$ and not just to $o(1/n)$. The peculiarity is that we shall find that the nature of this rule, so long as there is one, will not affect the results.

The basic expressions now lengthen to (we shall drop the γ subscripts throughout this discussion for clarity of formulas)

$$\vec{\text{ave}}\, t = \theta + L + \frac{1}{n}K + \frac{1}{n^2}M\,, \qquad \vec{\text{var}}\, t = \frac{1}{n}J + \frac{1}{n^2}N\,.$$

If our estimand has bias L_e, then the minimum of the asymptotic average square deviation, when a base of estimate is to be selected from a one-parameter family, will occur when

$$L = L_e - \left(K + \frac{1}{2}\frac{dJ}{dL}\right)\frac{1}{n}\,,$$

and the corresponding values of the asymptotic average square deviation will be, when $L = L_e$,

$$\frac{1}{n}J + \frac{1}{n^2}(K^2 + N)\,,$$

but will be less by

$$\frac{1}{n^2}\left(K + \frac{1}{2}\frac{dJ}{dL}\right)^2\,,$$

when L is offset as above. Thus the improvement due to offset is by a factor asymptotic to

$$\frac{\left(K + \frac{1}{2}\frac{dJ}{dL}\right)^2}{nJ}\,.$$

Notice that neither the formula for $L - L_e$, nor that for the fractional improvement in variance due to its use involves the coefficients M and N of $1/n^2$. Whatever modification of the definitions of $\vec{\text{ave}}$ and $\vec{\text{var}}$ is adopted, the leading terms in the amount of, and the profit from, optimum offset are the same.

B.3 Choice of Expression

If θ is replaced by θ^2, $\sqrt{\theta}$, or $\log\theta$, the discussion of the last section can be repeated. What effect will there be on asymptotic variances and biases?

Investigation shows ([8], Sec. 14) that the main changes are

1. An over-all scale change by a factor expressing the derivative of one mode of expressing the parameter with respect to the other,
2. The addition of a term in K_γ,
3. Non-linear effects for large biases.

These effects do not have an important influence on any of the uses to which we expect to put H_γ, J_γ, K_γ, L_γ, or Δ_γ. Consequently, we may take the particular mode of expression of the parameter as not being of major importance.

B.4 Estimates of Scaling Type

When the parameter θ is to be a function of a scaling parameter, λ, there are certain conveniences in the choice of a certain mode of expression. (And, indeed, it can be argued that, insofar as this choice has minor effects on the comparison of biases and variabilities, this convenient choice is also the most proper choice, so far as these minor changes go.) This choice is a logarithmic expression of scale, say in the form $\lambda = e^\tau$, where τ is now to play the role played by θ above.

What are some of these conveniences?

1. If $G_\gamma((z - z_0)/\lambda)$ is a scale pencil of distributions, and if we write $w = \log(z - z_0)$ and let $H(w, \tau)$ be the corresponding family of distributions of w parametrized by τ, then τ is a location parameter.
2. Most, and by a device all, of the most natural bases of scaling have equivalent forms in which the bias of estimating λ is independent of λ.

These equivalent forms include

$$\frac{1}{p}\log\frac{1}{n}\sum|z_i|^p \quad \log(p\%\text{-interfractile distance})\,,$$

and (here is a device)

$$\log\lambda_A\,, \qquad \text{where } \frac{1}{n}\sum e^{-b(z/\lambda_A)^2} = A\,,$$

the last being asymptotically equivalent, as a base of estimate, to the Gaussian mean $(1/n)\sum e^{-cz^2}$ where $c\lambda^2$, which is the asymptotic value of $c\lambda_A^2$, equals b.

This choice is recommended for all problems of scaling type, and was used in the treatment of scaling problems in contaminated distributions reported in Section 10.

B.5 Balancing Bias and Variance in the Example

Formulas for balancing bias and variance were attained in Section B.2, but no indication has so far been given that they may be practically usable. For a specific alternative pencil can never be identified in practice, where the alternative is always an unknown pencil, one of many possibilities. If the formula yields widely different choices of bases of estimate for

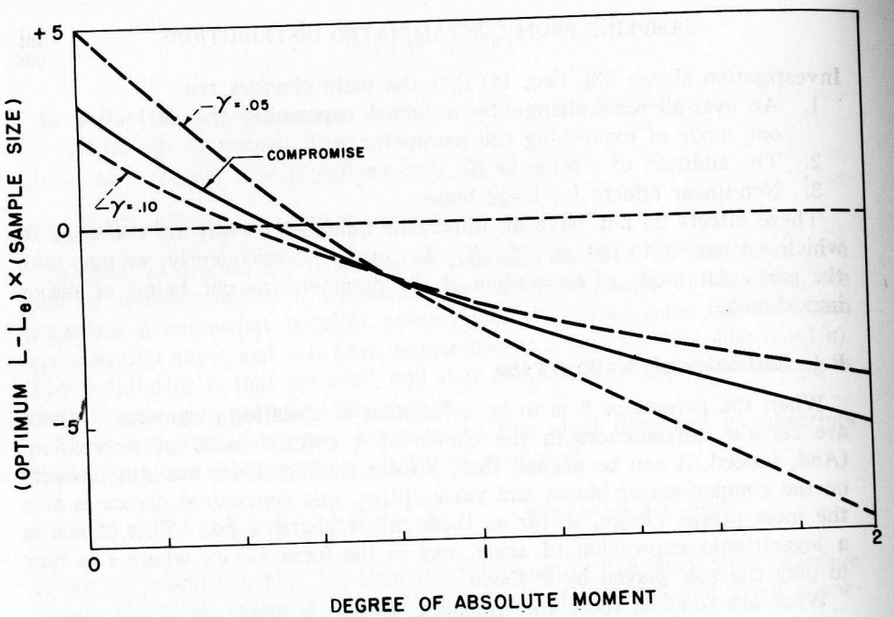

FIG. B-1. Optimum (for Specific γ) and Compromise Offsets in Bias

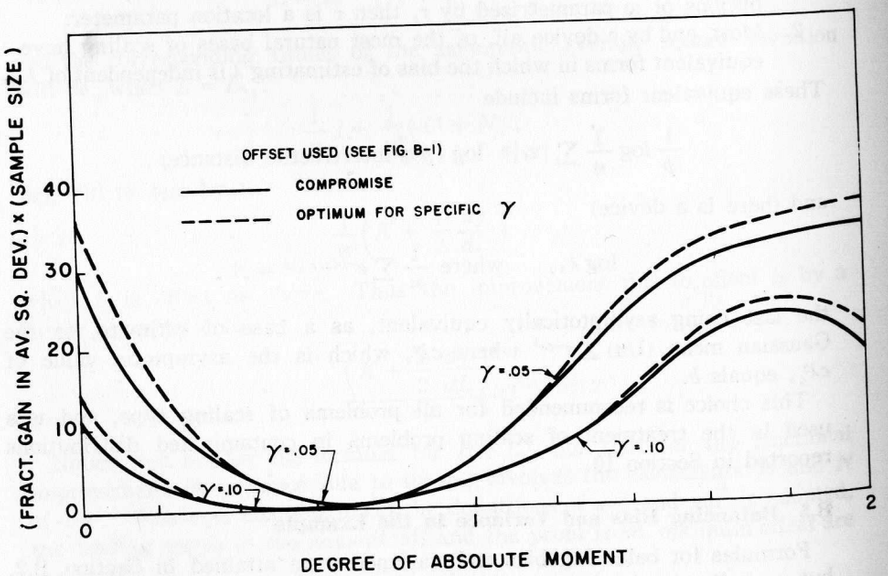

FIG. B-2. Fractional Gain in Average Square Deviation from Offset in Bias

possible alternative pencils, in our instances for different γ, it will be of little practical use.

A brief look at the scaling of contaminated distributions thus seems in order. The case of replacing a moderate absolute moment (one with $p < 2$) by a Gaussian mean of similar bias is a convenient example. Figure B-1 shows values of the offset in L as "(optimum $L - L_e$) × (sample size)" for $\gamma = .05$ and $\gamma = .10$ for $0 \leqq p \leqq 2$. The curve marked "compromise" is, in most places, the geometric mean of the curves for individual γ. Figure B-2 shows curves for "(fractional gain in average square deviation) × (sample size)" for the same situations. Because the loss in adopting a non-optional offset is a quadratic function, the compromise gives quite satisfactory results.

The general order of *fractional* gain is from $0/n$ to $30/n$. This would be negligible for samples of size 1000, so that these should presumably be treated as very large. Depending upon the degree of the moment, the gain for samples of size 100 would frequently be important. Samples of size 100 might thus frequently be wisely treated as moderately large.

This whole discussion is incomplete, because it has not taken into account the fact (see Fig. 9 in Section 12) that slightly different b values are required to match bias with a single absolute moment for $\gamma = .05$ and $\gamma = .10$. However, the indications (i) that useful offsets for several alternative families are possible, and (ii) that "moderately large" may mean approximately 100, are still useful and illuminating.

REFERENCES

[1] Cunningham, L. B. C., and Hynd, W. R. B. Random processes in problems of air warfare. *J. Roy. Stat. Soc.*, Ser. B, 1946, **8**, 62–85, disc. 85–97.

[2] Fraser, D. A. S. Generalized hit probabilities with a Gaussian target. *Ann. Math. Stat.*, 1951, **22**, 248–55; 1955, **24**, 288–94.

[3] Tukey, J. W. Sampling from contaminated distributions, preliminary report (abst.). *Ann. Math. Stat.*, 1946, **17**, 501.

[4] Tukey, J. W. Sampling from contaminated distributions, preliminary report (abst.). *Bull. Amer. Math. Soc.*, 1946, **52**, 828–29.

[5] Tukey, J. W. Sampling from contaminated distributions, 2. Scaling by and for power-means—the asymptotic case. Statistical Techniques Group, Memorandum Report No. 25, Princeton, N.J.: Princeton University, (1949 draft, unpublished).

[6] Harris, T. E., and Tukey, J. W. Sampling from contaminated distributions, 3. Measures of location and scale which are relatively insensitive to contamination. Statistical Research Group, Memorandum Report No. 31, Princeton, N.J.: Princeton University, July 1949, (duplicated).

[7] Tukey, J. W. Sampling from contaminated distributions, 4. The truncated mean in moderately large samples. Statistical Research Group, Memorandum Report No. 32, Princeton, N.J.: Princeton University, July 1949 (duplicated).

[8] Tukey, J. W. Sampling from contaminated distributions, 5. Scaling by and for percentiles and exponential averages. Statistical Research Group, Memorandum Report No. 33, Princeton, N.J.: Princeton University (1949 draft, unpublished).

[9] Tukey, J. W. Sampling from contaminated distributions, 6. Skeleton tables related to contaminated distributions at scale 3. Statistical Research Group, Memorandum Report No. 34, Princeton, N.J.: Princeton University, July 1949 (duplicated).

[10] Tukey, J. W. Estimation in the alternative family of distributions (abst.). *Proc. Int. Cong. Math.*, 1950, **1**, 586, 1952.
[11] Pearson, E. S. The analysis of variation in cases of non-normal variation. *Biometrika*, 1931, **23**, 114–33.
[12] Eden, T., and Yates, F. On the validity of Fisher's z-test when applied to an actual example of non-normal data. *J. Agric. Sci.*, 1933, **23**, 6–16.
[13] Pitman, E. J. G. Significance tests which may be applied to any population. *J. Roy. Stat. Soc.*, 1937 suppl., **4**, 119–30.
[14] Pitman, E. J. G. Significance tests which may be applied to samples from any population, II. The correlation coefficient test. *J. Roy. Stat. Soc.*, 1937 suppl., **4**, 225–32.
[15] Pitman, E. J. G. Significance tests which may be applied to samples from any population, III. The analysis of variance test. *Biometrika*, 1937, **29**, 322–35.
[16] Welch, B. L. On the z-test in randomized blocks and Latin squares. *Biometrika*, 1937, **29**, 21–52.
[17] Box, G. E. P., and Andersen, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumptions, *J. Roy. Stat. Soc.*, Ser. B, 1955, **17**, 1–26, disc. 26–34.
[18] Baker, G. A. Empiric investigation of a test for homogeneity for populations composed of normal distributions, *J. Amer. Stat. Assoc.*, 1958, **53**, 551–57.
[19] de Helguero, F. Per la risoluzione delle curve dimorphiche. *Biometrika*, 1905(-06), **4**, 230–31.
[20] de Helguero, F. Per la risoluzione delle curve dimorphiche, *Mem. R. Accad. dei Lincei* (*Rome*), Ser. 5, 1906, **6**, 163–203.
[21] Edgeworth, F. Y. On the representation of statistics by mathematical formulae, Pt. 2. *J. Roy. Stat. Soc.*, 1899, **62**, 125–40.
[22] Feller, W. On a general class of contagious distributions. *Ann. Math. Stat.*, 1943, **14**, 389–400.
[23] Gurland, J. Some interrelations among compound and generalized distributions. *Biometrika*, 1957, **44**, 265–68.
]24] Pearson, K. Contributions to the mathematical theory of evolution, 7. *Phil. Trans. Roy. Soc.* (*London*), Ser. A, 1894, **185**, 71–110.
[25] Pearson, K. On some applications of the theory of chance to racial differentiation. *Phil. Mag.*, 1901, **6**(1), 110–24.
[26] Pollard, H. On the relative stability of the median and arithmetic mean, with particular reference to certain frequency distributions which can be dissected into normal distributions. *Ann. Math. Stat.*, 1934, **5**, 227–62.
[27] Preston, E. J. A graphical method for the analysis of statistical distributions into two normal components. *Biometrika*, 1954, **40**, 460–64.
[28] Jeffreys, H. An alternative to the rejection of observations. *Proc. Roy. Soc.* (*London*), Ser. A, 1932, **137**, 78–87.
[29] Levi, B. On the form of composite frequency curves (Spanish). *Math. Notae*, 1951, **11**, 87–109.
[30] Medgyessy, P. Anwendungsmöglichkeiten der Analyse der Wahrscheinlichkeitsdichtefunktionen bei der Auswertung von Messungsergebnissen. *Z. angew. Math. u. Mech.*. 1957, **37**, 128–39.
[31] Navratil, J. Determination of the parameters of a compound normal distribution. *Pokroky Mat. Fys. Astr.*, 1958, **3**, 41–45.
[32] Putnam, C. R., and Wintner, A. On the addition of symmetric normal frequency curves. *Math. Notae*, 1951, **11**, 79–86.

[33] Teichroew, D. The mixture of normal distributions with different variances. *Ann. Math. Stat.* 1957, **28**, 510–12.
[34] Watanake, Y., Wajiki, I., and Kawashiro, T. On the compound normal distributions. *J. Gakugei Tokushima Univ. Nat. Sci. Math.*, 1956, **7**, 53–65.
[35] Wintner, A. Infinite divisible symmetric laws and normal stratifications. *Publ. Inst. Stat. Univ. Paris*, 1957, **6**, 326–36.
[36] Student. Errors of routine analysis. *Biometrika*, 1927, **19**, 151–69.
[37] Student (William Sealy Gosset). *"Student's" Collected Papers*. London: Biometrika Office, 1942.
[38] Tukey, J. W. Asymptotic moments and expectations (abst.). *Bull. Amer. Math. Soc.*, 1948, **54**, 644–45.
[39] Tukey, J. W. Asymptotic moments and expectations. Statistical Techniques Group, Memorandum Report No. 4, Princeton, N.J.: Princeton University, May 1949 (duplicated).
[40] Tukey, J. W. Asymptotic variances and average values. (Being revised for the *Annals of Mathematical Statistics*.)
[41] Hotelling, H. The place of statistics in the university. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, Calif.: Univ. California Press, 1949.
[42] Jung, J. On linear estimates defined by a continuous weight function. *Arkiv för Matematik*, 1955(1954–58), **3**(15), 199–209.
[43] Jeffreys, H. *Theory of Probability*, 2d ed. Oxford: Clarendon Press, 1948.
[44] Mandel, J., and Stiehler, R. O. Sensitivity—a criterion for the comparison of methods of test, *J. Research N.B.S.*, 1954, **53**, 155–59.
[45] Cramér, H. *Mathematical Methods of Statistics*, Princeton, N.J.: Princeton Univ. Press, 1946.
[46] Fisher, R. A. A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Not. Roy. Astron. Soc.*, **80**, 758–70.
[47] Fisher, R. A. *Contributions to Mathematical Statistics*, New York: Wiley, 1950.
[48] Burrau, Ø. The mean error as a measure of uncertainty (Danish). *Mat. Tidskr.*, Ser. B, 1943, 9-16 (cf. *Math. Rev.*, 1946, **7**, 130).
[49] Fisher, R. A. On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc.* (*London*), Ser. A, 1922, **222**, 309–68.