Second Order Approximations [1]

Willem Albers
University of Limburg
P.O. Box 616
6200 MD Maastricht

Before embarking on a discussion of second order approximations, it is perhaps a good idea to first say a few words about the need for such approximations. To begin with, we would of course prefer to use exact results only. However, it turns out that in statistics this is very often not feasible, which forces us to settle for some kind of approximation. Typically such an approximation is of an asymptotic nature. Consider for example the following situation: we are interested in some characteristic of a random quantity, such as the mean value of the performance of a particular medical treatment, or the variance of a new method for weighing items of a given type. To obtain the desired information, we draw a sample $X_1,...,X_n$, that is, we collect n independent measurements of the quantity under consideration. From these we evaluate an appropriate function $T_n = T_n(X_1,...,X_n)$ for our purpose, which could e.g. be testing or estimation. Then a result like:

$$T_n \text{ is } \mathrm{AN}(\mu_n, \sigma_n^2) \qquad (1)$$

where 'AN' stands for 'asymptotically normal' and μ_n and σ_n^2 are the known mean and variance of this normal distribution, will provide us with first order approximations to the quantities which are of interest for assessing the quality of the procedure we use. For an estimator this could be its variance and for a test its size and power (which stand for the probability of rejecting the null hypothesis when it is true and when it is not true, respectively). To be just slightly more precise, (1) can be stated equivalently as

1. This paper was presented at a meeting organized by the Mathematics Department of the University of Leiden on the occasion of the honorary degree in mathematics which was awarded by this University to Professor Erich L. Lehmann on February 8, 1985.

$$P(\frac{T_n - \mu_n}{\sigma_n} \leqslant x) = \Phi(x) + o(1), \tag{2}$$

where Φ stands for the standard normal distribution function. Establishing a second order approximation now entails replacing the '$o(1)$' in (2) by 'something better'. (You see that technicalities are ruthlessly estimated!) A first obvious advantage of this additional effort will be that the numerical approximations will typically improve.

The approach above is rather single-minded in the sense that one single statistic at a time is considered. Usually, we are interested not in a particular statistic, but in a particular statistical problem, for which as a rule several statistics present themselves as possibilities on which to base the solution. Moreover, quite often several of these candidates are first order equivalent. That is, the same result (2) (with the same μ_n and σ_n) holds for all of these statistics. At first sight, this looks delightful: as long as the good choices are equivalent, it does not matter much which one we pick. However, the same objection as above applies: the equivalence holds to first order only and the finite sample behaviour of the statistics may (and in fact often does) differ quite a bit.

It is to this problem that Professor LEHMANN, together with Professor HODGES, drew attention with admirable clarity in his 1970 paper in the *Annals of Mathematical Statistics* with the concise title 'Deficiency'. We shall now briefly review the main idea and some illustrative simple examples from this paper. Suppose we are given two statistical procedures A and B for a certain statistical problem. For simplicity of presentation, let us assume that it is known beforehand that A is the better of the two. For each $n = 1,2,\ldots$, we can determine the number k of observations which is needed by the poorer procedure B to match the performance (e.g. reach the same power or the same variance) of procedure A when based on n observations. Clearly $k \geqslant n$, and since it will depend on n, we shall denote it by k_n. Typically, people have been studying the behaviour of the ratio

$$e_n = \frac{n}{k_n} , \tag{3}$$

which is called the relative efficiency of B with respect to (wrt) A. Quite often it can be shown that e_n tends to a limit e, which is called the asymptotic relative efficiency (ARE) of B wrt A. Now HODGES and LEHMANN point out that it would be more natural to study the difference $k_n - n$, rather than the ratio in (3). They call this difference, which is nothing but the additional number of observations required by the poorer procedure, the deficiency d_n of B wrt A. And in analogy to the above, its limit d, if it exists, is called the asymptotic deficiency of B wrt A.

Note that this discussion is a bit deceptive in the following sense: it may be true that a difference is more natural to study than a ratio, but it is also true that it is more difficult to handle. In fact, first order results like (2) suffice to evaluate e_n and e, but for d_n and d second order approximations are required.

Consequently, before going on, it makes sense to figure out in which cases efficiencies suffice and in which cases the additional effort involved in obtaining deficiencies is worthwhile. A situation of the first kind arises when $e < 1$. Then it is already clear which of the two procedures is superior and information on d_n is at most useful to improve the numerical approximation. If $e=1$, however, the opposite situation occurs. Then the two procedures are first order equivalent and d_n (and possibly d) are vital in finding out which of the two is best, and by how much the poorer one falls short of the better one. Hence in what follows we shall concentrate on cases where $e=1$.

A first, very simple example, is the following. Let $X_1,...,X_n$ be a sample from an unknown distribution function F with mean ξ and variance σ^2. To estimate σ^2, two obvious estimators are available:

$$M_n=\frac{1}{n}\sum_{i=1}^{n}(X_i-\xi)^2 , \quad M_n'=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 , \tag{4}$$

where $\overline{X}=n^{-1}\sum_{i=1}^{n}X_i$. Clearly if ξ is known, we should use the first, and if ξ is unknown, we should use the second estimator. Quite often, however, we are somewhat 'in between': a value of ξ is available, but its reliability is not above all suspicion. Using M_n if the value of ξ is false can be disastrous; using M_n' if the value happens to be correct is a bit wasteful. Consequently, it is a robustness question we are facing: what is the price we may have to pay for the protection afforded by using M_n' rather than M_n?

This question can be nicely answered by using the deficiency concept. It turns out that the variances V_n and V_n' of the unbiased estimators in (4) equal

$$V_n=\sigma^4\frac{\gamma}{n} , \quad V_n'=\sigma^4\frac{\gamma(n-1)+2}{n(n-1)} , \tag{5}$$

where $\gamma=(\mu_4/\sigma^4)-1$, with μ_4 the fourth central moment of F. By elementary computations it follows from (5) that $e=1$ (hence we indeed have the interesting case of first order equivalence!) and moreover that d_n tends to a finite limit

$$d=\frac{2}{\gamma} . \tag{6}$$

If $F=\Phi$, we have that $\gamma=2$, and thus one additional observation is all we have to pay for the robustness in the normal case. This is not such a big surprise after all, since the estimators in (4) are in that particular case following χ_n^2- and χ_{n-1}^2-distributions, respectively. But (6) holds true in general. In practice F will quite often be slightly heavier-tailed than the normal, which results in $\gamma > 2$ and therefore in a value of d which is less than one. Incidentally, a possible interpretation of such a broken value of d can for example be provided by means of stochastic interpolation: use $n+1$ observations with probability d and n observations with probability $1-d$.

A second elegant example, also taken from the 'Deficiency' paper, is concerned with a related problem. Again $X_1,...,X_n$ is a sample from a distribution with mean ξ and variance σ^2, but here we assume the distribution to be normal. Instead of estimators we are now going to compare two test statistics. The testing problem involved is the classical $H_0:\xi=0$ against $H_1:\xi>0$. We can either use the test based on the sample mean $\overline{X}$ or Student's t-test. In the first case we reject if

$$\frac{n^{1/2}\overline{X}}{\sigma} \geqslant u_\alpha = \Phi^{-1}(1-\alpha)\,, \tag{7}$$

with α the size of the test. In the latter case we reject if

$$\frac{n^{1/2}\overline{X}}{(M_n')^{1/2}} \geqslant t_\alpha\,, \tag{8}$$

where M_n' is as in (4) and t_α is the upper α-point of the t-distribution. Note the analogy to the first example, the only difference being that the role of ξ and σ has been interchanged. Hence the issue now is whether we should rely on possible information about σ or use the t-test all the time. Again a nice answer is provided by the application of deficiencies. HODGES and LEHMANN demonstrate, using an elegant conditioning argument, that not only $e = 1$ but that the deficiency tends to the finite limit

$$d=\frac{u_\alpha^{\;2}}{2}\,, \tag{9}$$

with u_α as in (7). For $\alpha = 0.01, 0.025$ and 0.05 respectively, this leads to $d =$ 2.706, 1.921 and 1.353, respectively. Hence a very moderate number of additional observations suffices to achieve the desired robustness against deviations of σ. Of course these results are of an asymptotic nature, but it has also been demonstrated for this example that already for sample sizes as small as $n = 4$ and 8 a beautiful agreement with the exact values exists.

Summarizing the above, we can conclude that HODGES and LEHMANN have made a very nice point in their paper and have done so with great clarity. But this is by no means the end of the story. Even more important perhaps is the fact that their paper has stimulated a lot of further research. The authors encouraged this development by stating a number of questions at the end of their paper. Now it is well known that one fool may ask more than ten wise men can answer, but fortunately in this case two wise men have apparently succeeded in coming up with such questions that it has turned out worthwhile rather than foolish to try to answer them!

In the remainder of this paper we shall take a look at some of these questions and try to give an impression of what progress has been made in finding answers. The general background of the questions is more or less the following. As we have seen, the evaluation of deficiencies requires second order approximations. In a number of non-standard areas the derivation of such approximations presents serious technical difficulties. These have to be overcome before the often quite interesting application to deficiencies can be made.

The first question we consider is of the following nature. As we saw in (9), the asymptotic deficiency $d_{t,\overline{X}}$ of the t-test wrt the $\overline{X}$-test is finite and equals

$u_\alpha^2 / 2$. What happens if we do not stop at scale invariance but in addition require distribution-freeness? For example, would the asymptotic deficiency $d_{NSc,t}$ of the normal scores test, which is the best rank test for the normal case, wrt the t-test be finite? And if not, at what rate would the deficiency tend to infinity? Note that this is a 'good' question in the sense that $e_{NSc,t}$ is known to equal 1. This itself came as a bit of a surprise originally, as people at first used to think of rank tests as 'quick-and-dirty' methods, which probably sacrificed a lot of efficiency in exchange for their ease of application. Now that we know that this is not the case, at least to first order, we become eager for more and hope to show that the loss incurred is small to second order as well.

The second question is of a similar spirit: can rank tests be concocted which are second order equivalent wrt their parametric competitors, i.e. which have $d = 0$? Finally the third question we shall consider is about the relationship between tests and estimators. Suppose that for a problem two test statistics T_1 and T_2 are given and that each of these statistics gives rise to an estimator, say $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively, for the parameter of interest. Then it is well-known that typically the efficiency results for tests and estimators coincide, i.e. that

$$e_{T_1,T_2} = e_{\hat{\theta}_1,\hat{\theta}_2} . \tag{10}$$

The question then is whether a result like (10) also holds for deficiencies, that is, will $d_{T_1,T_2} = d_{\hat{\theta}_1,\hat{\theta}_2}$ also be true?

As a first step towards answering the first question, we shall indicate the nature of the difficulties that arise here. For a first order approximation we saw in (1) and (2) that an asymptotic normality result is needed. In the classical case of sums of independent random variables this is provided by the central limit theorem. The extension to second order is made by using Edgeworth expansions. To rank tests, and to distributionfree tests in general, these standard results cannot be applied. As is well known, quite a lot of effort has been devoted to obtaining asymptotic normality results for these cases. Hence it will come as no surprise that considerably more obstacles still have to be eliminated before second order approximations become available in this area. Here however we shall be content with noting that such approximations have indeed become available and we shall, ignoring all technicalities involved, concentrate on the results. At first sight, the result is discouraging: it turns out that the asymptotic deficiency $d_{NSc,t} = \infty$. However, it can in addition be shown that the deficiency satisfies

$$(d_{NSc,t})_n \sim \tfrac{1}{2} \log \log n , \tag{11}$$

which for all practical purposes is finite (and almost constant). Hence rank tests do live up to optimistic expectations to second order as well: the amount by which their performance falls short of that of their parametric counterparts is indeed enjoyably small. Incidentally, it is also possible to have a finite asymptotic deficiency. For example in the logistic case we have Wilcoxon's signed rank test as the optimal rank test and its asymptotic deficiency with respect to the optimal parametric test for the logistic case is indeed finite.

Next we have the related question about the possibility of rank tests with $d = 0$. In this connection we make the following reexamination of the foregoing. In the normal case we started out with the $\bar{X}$-test, which is the best parametric test. From there we went to the t-test, which is the best scale invariant test for the normal case. Then we decided to buy ourselves in addition distributionfreeness, and we moved on to the normal scores test, which, as noted before, is the best rank test for the situation at hand. Note now that this last step can be judged to be larger than strictly necessary. If we want the test to be distributionfree, we should perhaps look for the best distributionfree test, and not immediately restrict ourselves to rank tests, Indeed it turns out that such an intermediate possibility exists in the form of the best permutation test. The surprising result now is that the asymptotic deficiency $d_{P,t}$ of this best permutation test wrt the t-test satisfies

$$d_{P,t} = 0. \tag{12}$$

Hence once the price $d_{t,\bar{X}} = u_\alpha^2 / 2$ for scale invariance has been paid, no additional charge is involved to obtain distributionfreeness! Of course, besides forming some kind of answer to the second question, this result is more amusing than useful. In practice, people will probably be quite willing to pay the further price given in (11) to buy themselves in addition the ease of application of a rank test.

The answer to the third question simply seems to be yes. To give it a bit more body, we shall consider an illustrative example. Let S_n be the best rank statistic based on a sample $X_1,...,X_n$ from $F(x-\xi)$. Then introduce $S_n(\theta)$ which is the same statistic but now based on the shifted sample $X_1-\theta,...,X_n-\theta$. Let $\tilde{\theta}_n$ be such that $S_n(\tilde{\theta}_n) = E_{H_0}S_n$, then this is (ignoring once more the technical details) the so-called Hodges-Lehmann estimator of ξ. (This has nothing to do with the HODGES-LEHMANN 'Deficiency' paper, but it is simply difficult to discuss a contribution of Professor LEHMANN to statistics without running into other such contributions!) The relation between $\tilde{\theta}_n$ and the rank statistic is precisely the same as between the maximum likelihood estimator $\hat{\theta}_n$ and the parametric counterpart T_n of S_n. We already know that

$$e_{\tilde{\theta}_n,\hat{\theta}_n} = e_{S_n,T_n} = 1. \tag{13}$$

It turns out to be possible to show that the deficiency of $\tilde{\theta}_n$ wrt $\hat{\theta}_n$ (when properly defined) agrees to first order with the deficiency of S_n wrt T_n evaluated at size $\alpha = ½$. Hence if the limits involved exist, we indeed have that

$$d_{\tilde{\theta}_n,\hat{\theta}_n} = d_{S_n,T_n}. \tag{14}$$

In the above we have only considered one-sample results for rank tests. Similar results have also been obtained for the two-sample case and for simple linear rank statistics. Moreover, many contributions have been made to other areas in nonparametrics as well, L-statistics for example. Nevertheless, it is hoped that the brief sketch above has been sufficient to give an idea of the impact of Professor LEHMANN's work on a lot of recent research in statistics.