Charles Stein: a very personal view

Summary

During a long, distinguished career Charles Stein published a relatively small number of papers. Some are systematic and thorough, while others contain the germ of original ideas that stimulated future research, and a few suggested a program for future research. In this talk I will discuss some personal favorites (but avoid the two subjects for which he is best known, which will be represented elsewhere in these workshops). Included will be (i) his 1949 abstract on fixed width confidence intervals for a normal mean, (ii) his 1956 paper posing the problem of adaptive estimation and testing, (iii) the Efron-Stein inequality, and his beautiful example criticizing the likelihood principle.

Before 1950: the Problem

Let $X_1, X_2, \ldots$ be independent and normally distributed with unknown mean μ and variance $\sigma^2 > 0$. Given α, we desire to find a confidence interval I of prescribed width $2w > 0$ so that $\mathrm{P}\{\mu \in I\} \geq 1 - \alpha$ for all μ, σ^2.

Let $\bar{X}_n$ denote the average of $X_1, \ldots, X_n$ and $S_n^2 = (n-1)^{-1} \sum_1^n (X_i - \bar{X}_n)^2$.

Stein (1949)

Let z_α denote the $(1-\alpha/2)$ percentile of the standard normal distribution. Observe that the smallest integer exceeding $n_0 = (\sigma z_\alpha / w)^2$ is the sample size that would be required to obtain a $1-\alpha$ confidence interval of width $2w$ for a normal mean when σ is known. In 1949 Stein studied the asymptotic behavior of the procedure where one samples sequentially until it *appears* that the sample size n_0 has been achieved. More precisely, let

$$T = \min\{n : n \geq 4, (n-1)S_n^2 \leq a_n n^2 (w/z_\alpha)^2.$$

where $0 < a_n = 1 - a_0/n + o(1/n)$. Stein claimed that by choosing $a_0 = 3 - \rho + (1+z_\alpha^2)/2$, one would have as $w/\sigma \to 0$

$$\mathrm{P}\{|\bar{X}_T - \mu| \leq w\} = 1 - \alpha + 0 \cdot n_0^{-1} + o(n_0^{-1}) \tag{1}$$

and

$$\mathrm{E}(T) = n_0 + (1+z_\alpha^2)/2 + o(1). \tag{2}$$

Anscombe (1953), Chow and Robbins (1965)

Anscombe (1953), apparently in ignorance of Stein's 1949 abstract, discussed the same problem. Among a number of related results he gave a heuristic argument identifying the constant ρ, which he computed by solving a Wiener-Hopf integral equation numerically.

Chow and Robbins (1965) dropped the hypothesis that the X's are normally distributed and obtained the same first order asymptotics as Stein (1949) and Anscombe (1953). A flood of papers followed when it was realized that by considering only first order asymptotic results, one could apply the same ideas to a variety of problems.

Woodroofe (1977)

Woodroofe (1977) gave a mathematically satisfactory treatment of the Stein/Anscombe results described above. The tool that allowed him to give a rigorous evaluation of the constant ρ is what has come to be called nonlinear renewal theory developed largely for other purposes by Woodroofe (1976a,b) and Lai and Siegmund (1977, 1979).

Adaptive Estimation and Testing

In the Third Berkeley Symposium, 1956, Stein published two remarkable papers, one of which posed the problem of adaptive estimation and testing and stimulated an enormous amount of subsequent research by Bickel, Wellner, Ritov, Stone, and van der Vaart, among others. The problem is to provide a fully efficient estimator or test statistic for a finite dimensional parameter in the presence of an infinite dimensional nuisance parameter. In his paper Stein begins with a matrix equation that shows when adaptive estimation occurs in the parametric case, and he gives an asymptotic computation that indicates how one can obtain an asymptotically efficient test of the center of symmetry of an unknown symmetric distribution. Since then new, related research was produced regulary for about 40 years.

The Matrix Equation

Let J be an information matrix, partitioned into 3×3 blocks, $J_{i,j}$. Assume the parameter θ is similarly partitioned, so θ_1 is the parameter of interest, while (θ_2, θ_3) is a nuisance parameter. In general estimating the nuisance parameter (θ_2, θ_3) increases the asymptotic variance of $\hat{\theta_1}$ over what it would be if they were known. Stein shows that estimating θ_3 does not affect the asymptotic variance if $\hat{\theta_1}$ if $J_{1,3} = J_{1,2}J_{2,2}^{-1}J_{2,3}$. His proof is algebraic. A simple statistical proof is the observation that Stein's condition is the same as $\mathrm{E}[(\ell_1 - \mathrm{J}_{1,2}\mathrm{J}_{2,2}^{-1}\ell_2)\ell_3'] = 0$, i.e., the residual from regressing the score for θ_1 on the score for θ_2 is uncorrelated with the score for θ_3.

Relation to Fang and Siegmund (2019?)

Suppose that $dY_s = (\mu(s) + \xi f(s-t))ds + \rho Y_s ds + \sigma dW_s$. Here f is a positive kernel, for example, (i) the indicator that $s > 0$, (ii) the indicator of the interval $(0, \Delta]$, (iii) a symmetric probability density function centered at 0 with (unknown) scale τ, or (iv) the positive part function, s^+. The process is observed for $s \in T$. Here we assume that σ is known and equals 1.

The parameters of primary interest are t, ξ, which define a local signal. Let θ denote the nuisance parameters μ, ρ. Given t, the efficient score for testing $\xi = 0$ is

$$\frac{\partial \ell}{\partial \xi}(0, \hat{\theta}), \tag{3}$$

where $\hat{\theta}$ are maximum likelihood estimators of θ under the assumption that $\xi = 0$.

Asymptotic Analysis

By standard likelihood theory (3 is asymptotically distributed as

$$V_t = \frac{\partial \ell}{\partial \xi} - I_{\xi,\theta} I_{\theta,\theta}^{-1} \frac{\partial \ell}{\partial \theta}, \tag{4}$$

where I is the Fisher information matrix, partitioned according to the coordinates ξ, θ, and all expressions are evaluated at t, $\xi = 0$ and true values of θ.

We discovered by direct calculation and then realized using Stein's matrix equation that when $\xi = 0$ and T is large, estimation of ρ does not change the asymptotic variance of V_t.

The Efron-Stein Inequality

Let $X_1, X_2, \ldots, X_n$ be independent random variables and Let $Z = f(X_1, \ldots, X_n)$. Let $\mathrm{E_i(Z) = E(Z|X_j, j \leq i)}$ and $\mathrm{E^{(i)}(Z) = E(Z|Z_j, j \neq i)}$. Motivated by the Quenouille-Tukey jackknife, Efron and Stein proved that if $\mathrm{E(Z^2) < \infty}$, then $\mathrm{Var(Z) \leq \sum_1^n E[(Z - E^{(i)}(Z))^2]}$. Their proof relied on an ANOVA decomposition of Z into main effects and interactions of all orders. A very simple proof is provided by Boucheron, Lugosi, and Massart (2013), one of which involves a simple manipulation of the martingale $\mathrm{\sum[E_i(Z) - E_{i-1}(Z)]}$ and the observation that $\mathrm{E_{i-1}(Z) = E_i[E^{(i)}(Z)]}$. Boucheron, Lugosi, and Massart (2013) also describe numerous applications, including Poincaré inequalities, functions with bounded differences, and Rademacher averages.

The Likelihood Principle

In 1974 JRSSB Stein published a short note containing an example designed to illustrate a serious flaw in the likelihood principle as the basis for all statistical inference. After presenting his simple example, he confesses that it was derived from consideration of the following problem of sequential sampling. Let $X_1, X_2, \ldots$ be independent normally distributed random variables with mean μ and unit variance. Let $S_n = \sum_1^n X_k$ and for $b > 0$, let $T = \inf\{n : |S_n| \geq bn^{1/2}\}$. The log likelihood function is $\mu S_T - T\mu^2/2$, with maximum $S_T^2/T \approx b^2$, so the likelihood ratio statistic for testing $\mu = 0$ contains essentially no information about the value of μ.

Fixed Precision Confdidence Intervals Revisited

However, this stopping rule arises in another context that takes us full circle back to Stein (1949). We assume $\mu \neq 0$ and ask for the value of n such that the length of a confidence interval for μ is no more than $\epsilon|\mu|$. The appropriate inequality would be achieved by taking $n_0 \geq (z/\epsilon\mu)^2$ for a suitable quantile z from the normal distribution, but we do not know μ. In the spirit of Stein (1949) we approximate n_0 by the stopping rule $T = \inf\{n : |S_n| \geq (z/\epsilon)n^{1/2}$, use Anscombe's theorm to justify the asymptotic normality of $(S_T - T\mu)/T^{1/2}$, and obtain (approximately) the desired confidence interval, provided ϵ is small and μ is not too small.

More

In the context of repeated significance tests for clinical trials (Armitage, 1967), which are based on the stopping rule $T_m = \min(T, m)$, Siegmund (1977) suggested a method (also violating the likelihood principle) for obtaining confidence intervals that do not break down when μ is small, although they require a different calculation that is not based on the asymptotic normality of S_T/T. We order outcomes T_m, S_{T_m} in a counter clockwise fashion as we move around the boundary that defines T_m and compute tail probabilities according to that ordering. For example, for $b = 3,\ T = 100,\ S_T = bT^{1/2}$, a 95% lower confidence bound based on the asymptotic normality of S_T is 0.14. The bound based on finding μ that makes $\mathcal{P}\{T_m \leq 100,\ S_{T_m} > 0\} \approx 0.05$ is 0.07.

A Few Random Remarks I

The basis for Stein's approach to the central limit theorem, that for a standard normally distributed random variable Z, $\mathrm{E}(\mathrm{Zf}(\mathrm{Z}) - \mathrm{f}'(\mathrm{Z})) = 0$, is closely related to his eventual approach to discussing inadmissibility of of the maximum likelihood estimator of a multivariate normal mean, but this approach to inadmissibility does not appear in his earliest work on the subject. The analogous identity for Poisson approximation, $\mathrm{E}(\mathrm{Zg}(\mathrm{Z}) - \lambda \mathrm{g}(\mathrm{Z}+1))$ is also closely related to statistical ideas (cf. Robbins' 1956 paper introducing the Empirical Bayes method).

A Few Random Remarks II

Stein's IMS monograph, "Approximate Computation of Expectations," in addition to his elegant treatment of the normal and Poisson limit theorems, contains two related discussions in Chapters 5 and 6, which, as far as I know, have never been fully developed. These are an approach to (i) the probability of large deviations and to (ii) the local limit theorem, both for sums of independent random variables with probability density functions. Both problems have been thoroughly treated in the literature, using characteristic functions, a tool that Stein seems to have disliked.

A problem that seems to have occupied Stein's thoughts without a definitive solution is to find an approximate confidence interval for a multivariate normal mean. He made a number of proposals in 1962, but does not seem to have favored any of them. He returned to this problem in short section at the end of his 1981 paper with another proposal and references to related research of others.