IMS Collections
Vol. 0 (2009) 11–17

arXiv: math.PR/0000009

Some History of Optimality

Erich L. Lehmann

University of California, Berkeley

Contents

1. Combination of Observations

The earliest optimality considerations appear to be those of Laplace and Gauss at the beginning of the 19^{th} Century concerned with determining the best linear estimates of parameters in linear structures. Laplace calls these optimal estimates "the most advantageous" while Gauss refers to them as "the most plausible values." Various aspects of this problem were discussed throughout the 19^{th} Century under the heading "Combination of Observations." The version of the principal result generally accepted today is the so-called Gauss-Markov theorem. It states (in modern language) that the least squares estimates are the linear unbiased estimates with minimum variance. While restricted to linear estimates, the result is nonparametric in that it makes no assumptions about the distribution of the errors. For an account of this work see, for example, Stigler (1986), Hald (1998), or Chatterjee (2003).

2. Maximum Likelihood Estimation

Optimality next played an important part in Fisher's fundamental paper of 1922. In this paper (followed by a clarifying paper in 1925), Fisher considers estimation in parametric models, proposes the maximum likelihood estimator (MLE) as a generally applicable solution, and claims (but does not prove) that such estimators are consistent and asymptotically efficient efficient (i.e., that they minimize the asymptotic variance). Note that unlike the Gauss-Markov theorem, maximum likelihood estimation does assume that the distribution of the variables belongs to a given parametric family.

Maximum likelihood has become the most widely used method of estimation, and there has been an enormous amount of work connected with Fisher's claims concerning it. It has led to the discovery of superefficiency on the one hand and of second order efficiency on the other. Many counterexamples have been found (even to consistency), and a number of modifications (bias correction and replacement of the MLE by a consistent root of the likelihood equation, for example) have been proposed. The situation is complex, but under suitable restrictions Fisher's conjecture is essentially correct when the likelihood equation has a unique root. For a more precise statement, see Shao (1999), and for further discussion, for example, Efron (1982), Le Cam (1990), and Barndorff-Nielsen and Cox (1994).

3. The Neyman-Pearson Program

Least squares and maximum likelihood were first proposed on intuitive grounds and then justified by showing that they possessed certain optimality properties. Optimality as a deliberate program for determining good procedures was introduced in 1933 by Neyman and Pearson in a paper (on testing rather than estimation) appropriately called, "On the problem of the most efficient tests of statistical hypotheses." As they explain in the introduction:

> The main purpose of the present paper is to find a general method of determining tests which ... would be most efficient"[in the sense of minimizing the probability of erroneous conclusions].

In a certain sense this is the true start of optimality theory.

4. The Neyman-Pearson Theory of Hypothesis Testing

Neyman and Pearson (1933) implemented the above program by seeking, for any given situation, the test which, among all those controlling the probability of false rejection at a given level α, has the maximum power (and hence the minimum probability of false acceptance).

For testing a simple hypothesis against a simple alternative, they found the solution to this problem to be the likelihood ratio test. This result, which is mathematically quite elementary but has crucial statistical consequences, is known as the Neyman-Pearson Lemma.

It turns out that in some (very rare) cases the same test is most powerful against all alternatives under consideration. Such a uniformly most powerful (UMP) test is then the optimal solution to the given testing problem. Where a UMP test does not exist, additional criteria must be invoked.

For example, when nuisance parameters are present, Neyman and Pearson require that under the hypothesis the rejection probability be α for all values of the nuisance parameters. They call such rejection regions similar regions. As an important example, they show that the one-sided t-test is UMP among all similar regions.

For two-sided alternatives, one would not expect UMP tests to exist even without nuisance parameters. For such cases, Neyman and Pearson then impose the additional condition of unbiasedness, i.e., that the power of the test is $\geq \alpha$ for all alternatives. In follow-ups to their 1933 paper (1936 and 1938), they show, for example, that the two-sided t-test is UMP among all unbiased tests.

UMP similar or unbiased tests exist for important classes of testing problems concerning a single real-valued parameter (in the presence or not of nuisance parameters) but not for hypotheses such as

$$H : \theta_1 = \cdots = \theta_s$$

concerning several parameters.

A different condition, the principle of invariance (suggested by Hunt and Stein, unpublished), is successful in a number of important such multiparameter situations. If both the hypothesis and the class of alternatives remain invariant under a group G of transformations of the sample space, there does in these cases exist a UMP test among all tests invariant under G.

5. Wald's Optimality Criteria

A quite-different approach to optimality was initiated by Wald in his 1939 paper, "Contributions to the theory of statistical estimation and testing hypotheses," and was then developed further in a series of publications culminating in his 1950 book, "Statistical Decision Functions."

This approach was part of Wald's formulation of a general theory of decision procedures. Instead of seeking procedures that are uniformly optimal among some suitably restricted class of decision procedures, Wald proposes to minimize some global feature of the performance.

Specialized to hypothesis testing, these proposals reduce to:

i Maximize the average power, averaged with respect to some suitable weight function over the alternatives. For obvious reasons, Wald called such maximizing procedures Bayes solutions.

ii Maximize the minimum power over alternatives bounded away from the hypothesis.

6. The Hunt-Stein Theorem

An important connection between the earlier invariance approach and that of Wald is the Hunt-Stein theorem. It states that if a UMP invariant test exists under a group G satisfying certain conditions (called amenability), then this test also maximizes the minimum power over any invariant set of alternatives.[1]

To illustrate this theorem, consider the univariate linear hypothesis. Under the combined action of the groups of location, scale and orthogonal transformation, a UMP invariant test exists. Since these groups satisfy the Hunt-Stein conditions, the resulting test therefore maximizes the minimum power against all invariant sets of alternatives.

As a second example, consider the multivariate one-sample problem. Hotelling's T^2-test is UMP among all tests that are invariant under the group G of all non-singular linear transformations. Since G is not amenable, the Hunt-Stein theorem does not apply. The resulting maximin problem poses considerable difficulties.

[1] The 1946 paper by Hunt and Stein, "Most stringent tests of statistical hypotheses," containing this theorem was never published. The theorem appeared in print for the first time in Lehmann, "Testing Statistical Hypotheses," John Wiley, 1959.

7. Some Extension of the Neyman-Pearson Theory

The Neyman-Pearson theory has been extended in a number of directions. The following are two extensions of the Neyman-Pearson Lemma, which is so basic to this theory.

(i) Sequential Analysis

During World War II, Wald proposed sequential procedures, as a way of obtaining good power with fewer observations. In particular, he suggested the probability ratio test for testing a simple hypothesis against a simple alternative. This test continues observation as long as the likelihood ratio remains between two fixed limits and takes the indicated decision (acceptance or rejection) as soon as it falls outside these limits.

In 1948, Wald and Wolfowitz proved the remarkable result that for testing a simple hypothesis against a simple alternative, the sequential probability ratio test, among all tests with the same (or smaller) probabilities of error, minimizes the expected number of observations both under the hypothesis and the alternative.

(ii) Robust Inference

All the work reported so far (except for the Gauss-Markov theorem) was carried out under the assumption of an underlying parametric model. In practice, such an assumption can be expected to hold at best approximately. As a more realistic formulation, Huber (1964) suggested replacing the assumption of a parametric model by that of a neighborhood of such a model.

In the following year, he obtained the analog of the Neyman-Pearson Lemma. For testing the neighborhood of a distribution P_0, the test maximizing the minimum power over the neighborhood of an alternative P_1 is a censored version of the likelihood ratio test of P_0 against P_1.

(iii) Multiple Testing

A very different extension of the Neyman-Pearson theory that is of great practical importance deals with the situation in which a number of hypotheses (sometimes a very large number) are being tested rather than just one. Unlike the work discussed so far, which is classical, the theory of multiple testing is an area of very active ongoing research.

The first problem here (before optimization) is to find a suitable generalization of the concept of significance level that would provide satisfactory control of the probability of false rejections. After this, maximin tests have been obtained, which require, however, not only unbiasedness and invariance but also a condition of monotonicity. Surveys of the current state of this work are provided by Shaffer (2004, 2006).

8. Large Sample Optimality of Testing

A small-sample theory of optimal estimation parallels that of optimal testing sketched in Sections 4-7, with concepts such as unbiasedness, equivariance (instead of invariance), and minimax variance (instead of maximin power), and will not be

discussed here. Asymptotic optimality for estimation goes back to Fisher (1922), as mentioned in Section 2, and is defined as minimum asymptotic variance.

For testing, asymptotic optimality is considerably more complex, both conceptually and technically. It was first studied by Wald in 1941. Consider testing a simple hypothesis $\theta = \theta_0$ against a simple alternative $\theta = \theta_1$. If we keep both θ_0 and θ_1 fixed, and carry out the tests at a fixed level α, the power of any reasonable test sequence will tend to 1. Thus any such test sequence will in a trivial sense be asymptotically UMP.

A more useful approach is obtained by considering a sequence of alternatives

$$\theta_n = \theta_0 + h/\sqrt{n} \tag{8.1}$$

For a sequence with fixed h, the power will typically tend to a limit between 0 and 1 as $n \to \infty$. As h varies from 0 to ∞, the limiting power will be an increasing function of h, going from α to 1; we shall call this the asymptotic power function. A sequence of tests can then be defined as being asymptotically most powerful (AUMP) if it maximizes the asymptotic power for all h.

Unlike the finite sample situation where UMP tests exist only rarely and then are unique, it turns out that AUMP tests exist under very weak assumptions, and that in fact many different AUMP tests exist for the same situation, among them the likelihood ratio test, the Wald test, and the locally most powerful (Rao) test. To distinguish them, one must resort to higher order asymptotics. (See, for example, Barndorff-Nielsen and Cox (1994)). An exposition of the first order theory due to Le Cam can be found in Lehmann and Romano (2005).

9. Optimal Design

In generalization of minimizing the variance of an estimator, optimal design theory is concerned with determining the design that minimizes some function of the covariance matrix of the best linear estimators of the parameters in question. In particular, D-optimality minimizes the determinant, E-optimality is a minimax criterion, and so on. After some isolated earlier efforts, the problem of optimal design was studied systematically by Kiefer in more than 40 papers between 1958 and his early death in 1981. They make up vol. III of his collected papers.

10. A Culture Clash

Not everyone was enthusiastic about optimality as a guiding principle. Criticism was highlighted at a 1958 meeting of the Royal Statistical Society at which Kiefer presented a survey talk on "Optimum Experimental Designs." It was a discussion paper, and the reaction was nearly uniformly negative. The core of the disagreement is stated clearly when Barnard quotes Kiefer as saying of procedures proposed in a paper by Box and Wilson (1951) that they "often [are] not even well-defined rules of operation." Barnard's reply:

> In the field of practical human activity, rules of operation which are not well defined may be preferable to rules which are.

The conflict is discussed by Henry Wynn in his introduction to a reprint of Kiefers paper in "Breakthroughs in Statistics," vol. I (Kotz and Johnson, Eds.). Wynn calls it "a clash of statistical cultures."

This clash is between the Wald school of abstraction and optimality on the one hand and what Tocher, in his discussion of Kiefer's paper, calls "the school of British experimental design – a very practical people," on the other.

11. Tukey's Criticism

Criticism of optimality was not confined to England. An outspoken American critic questioning the value and importance of optimization was John Tukey. His attitude is indicated by the titles of two philosophical papers in 1961 and 1962, which are titled respectively, "The tyranny of the best" and "Dangers of optimization."

Tukey's concern with optimality had its origin in the fact that at the time optimization had become the dominant interest of mathematical statistics. In the 1962 paper, he writes:

> Some [statisticians] seem to equate [optimization] to statistics an attitude which, if widely adopted, is guaranteed to produce a dried-up, encysted field with little chance of real growth.

12. Conclusion

That optimality considerations had become so dominant is explained by the historical situation. Periods of great innovation are followed by periods of consolidation, in which the new work is given a final (or, as Tukey says, encysted) form. Such a dichotomy is also discussed by Huber (1975). Thus, the revolutionary work of Student and Fisher was followed by the optimization approach of Neyman, Pearson, Wald and their students described in this paper.

Today we are faced with the opposite situation. We are confronted with new problems arising from immense data sets and involving problems of great complexity. Ad hoc solutions are proposed and tried out on a few examples. This is a natural first step, but eventually we will want to justify the solutions at which we have arrived intuitively and by trial and error. A theoretical underpinning will be provided and the conditions will be found under which these solutions are the best possible.

References

[1] Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics.* Chapman & Hall, London.

[2] Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society (B)*, **13**, 1–45.

[3] Chatterjee, S. K. (2003). *Statistical Thought.* Oxford University Press.

[4] Efron, B. (1982). Maximum likelihood and decision theory. *Annals of Statistics,* **10**, 309–368.

[5] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London (A)*, **222**, 401–415.

[6] Fisher, R. A. (1925). Theory of statistical estimation. *Cambridge Philos. Soc.*, **22**, 700–725.

[7] Hald, A. (1998). *A History of Mathematical Statistics - from 1750 to 1930.* John Wiley, New York.

[8] Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73–101.

[9] Huber, P. (1965). A robust version of the probability ratio test. *Annals of Mathematical Statistics*, **36**, 1753–1758.

[10] Huber, P. (1975). Applications vs. abstraction: the selling out of mathematical statistics? In *Proc. Conference on Directions of Mathematical Statistics. Suppl. Adv. Prob.*, **7**, 84–89.

[11] Kiefer, J. C. (1959). Optimum experimental designs (with discussion). *Journal of the Royal Statistical Society (B)*, **21**,273–319. (Reprinted in Kotz and Johnson (1992). *Breakthroughs in Statistics, Vol. 1.*, Springer-Verlag.

[12] Kimball, G. E. (1958). A critique of operations research. *J. Wash. Acad. Sci.*, **48**, 33–37.
[13] Kotz, S. and Johnson, N. L. (Eds. 1992, 1997). *Breakthroughs in Statistics*, **1**. Springer-Verlag, New York.
[14] Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. of Calif. Publ. in Statist.*, **1**, 277–330.
[15] Le Cam, L. (1990). Maximum likelihood an introduction. *ISI Review*, **58**, 153–171.
[16] Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses* (3rd Ed.). Springer, New York.
[17] Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. (A)*, **231**, 289–337.
[18] Neyman, J. and Pearson, E. S. (1936, 1938). Contributions to the theory of testing statistical hypotheses. *Statist. Res. Memoirs*, **1**, 1–37; **2**, 25–57..
[19] Pearson, E. S. (1939). "Student" as a statistician. *Biometrika*, **30**, 210–250.
[20] Shaffer, J. P. (2004). Optimality results in multiple hypothesis testing. In *Proceedings of the First Lehmann Symposium.* J. Rojo and V. Pérez-Abreu (Eds). IMS LNMS, **44**, 11–35.
[21] Shaffer, J. P. (2006). Recent developments towards optimality in multiple hypothesis testing. In *Proceedings of the Second Lehmann Symposium.* J. Rojo (Ed). IMS LNMS, **49**, 16–32.
[22] Shao, J. (1999). *Mathematical Statistics.* Springer, New York.
[23] Stigler, S. (1986). *The History of Statistics.* Harvard University Press, Cambridge, MA.
[24] Tukey, J. W. (1961). Statistical and quantitative methodology. In *Trends in Social Science*, (D. P. Ray, Ed.). Philosophical Library, New York. (Reprinted in Vol. III of Tukey's Collected Works.)
[25] Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics.* Reprinted in Vol. III of Tukey's Collected Works.
[26] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.*, **54**, 426–482.
[27] Wald, A. (1950). *Statistical Decision Functions.* John Wiley, New York.
[28] Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, **19**, 326–339.