

Some of Sam Greenhouse's contributions to statistical methods[‡]

Mitchell H. Gail^{*,†}

*Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute,
6120 Executive Boulevard, Room EPS 8032, Bethesda, MD 20892-7244, U.S.A.*

SUMMARY

I briefly survey some areas of Sam Greenhouse's contributions to statistical methods before focusing on three examples. These examples illustrate Sam's ability to identify problems of practical importance and to make valuable contributions to their solutions. Many of his papers continue to provide important guidance and insight because he dealt with issues of enduring practical importance. Published in 2003 by John Wiley & Sons, Ltd.

1. INTRODUCTION

I had the good fortune to take Professor Samuel W. Greenhouse's (Sam's) course in multivariate analysis at George Washington University. He led us through the intricacies of that distribution theory with great enthusiasm. There was no doubt that Sam enjoyed theory, even if the students did not always fully appreciate it. But Sam's contributions to theory and statistical methods were rooted in applications related to his consulting responsibilities. He described his work environment and the role of theory in an article on the National Institutes of Health [1]: 'One thing was not subject to any debate, namely, we were at the NIH, in accordance with Harold Dorn's directive, in order to provide the best statistical advice to questions posed to us by Intramural scientists... A secondary objective was, research in methodology and theory.'

It is difficult to summarize Sam's many contributions to statistical methods in a short article. I have chosen three examples that illustrate his role at the interface of theory and application. But, before turning to these examples, I list some of the other areas of statistical methodology to which he contributed.

Collaborating with Nathan Mantel and Abraham Goldin, Sam developed methods of bioassay to study the effects of various doses of antileukaemic agents in mice. Those innovative analyses allowed for more than one antileukaemic agent [2], a balancing of toxic

*Correspondence to: Mitchell Gail, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Boulevard, Room EPS 8032, Bethesda, MD 20892-7244, U.S.A.

†E-mail: gailm@exchange.nih.gov

‡This article is US Government work and is in the public domain in the U.S.A.

with therapeutic effects [3], and an investigation of the timing of 'citrovorum rescue,' in which aminopterin kills the leukaemic cells and folic acid (citrovorum factor) is later used to protect the host [4].

A citation classic by Greenhouse and Geisser [5] provides methods of analysis for profile data, which characterize an individual's pattern of multivariate measurements. They showed how to use standard analysis of variance (rather than more complex multivariate analysis) to analyse profile data. They presented modifications of the distribution theory appropriate to the simpler analysis and defined settings where standard distribution theory for the simpler approach is exact. Sam's other work on multivariate methods includes papers with Max Halperin on multiple comparisons [6, 7], such as a method to compare adjusted group means estimated from an analysis of covariance [7].

Sam played a prominent role as a collaborator and advisor on clinical trials. In view of the complex aims and conduct of real trials, he discussed the limitations of over-simplified formal hypothesis testing as the basis of interpretation [8]. Also, he contributed to the theory of adaptive treatment allocation, designed to reduce the number of subjects assigned to the inferior treatment [9] and to methods for restricted sequential designs [10, 11].

Among Sam's other contributions to methods were a paper on the validity of matched and unmatched cohort and case-control studies [12], work with Nathan Mantel on continuity-correction [13] and on the equivalence of maximum-likelihood and moment estimators in probit analysis [14], and work with Joseph Gastwirth to apply biostatistical methods to legal issues. An application to employment discrimination motivated work on methods for estimating a common relative risk, rather than relative odds [15], and an extension of the Cornfield inequality to bound the effects of unmeasured covariates was also applied to issues of employment discrimination [16].

Although not exhaustive, this brief survey indicates the breadth of Sam's contributions to methods and associated applications. We now treat three examples in greater detail.

2. EVALUATION OF DIAGNOSTIC TESTS

Shortly after joining Dr Harold Dorn's statistical unit in the Division of Public Health Methods in 1947, the unit (which also included Jerry Cornfield, Nathan Mantel, Jack Lieberman, and George Deal) was transferred to the National Cancer Institute [1]. Presumably it was Sam Greenhouse's turn to answer the consultants' phone when Dr John Dunn called for statistical advice on the evaluation of diagnostic tests for cancer [17, 18]. This consultation led to a paper that is still widely cited and used. I will recast the notation and terminology in today's parlance, but the concepts and formulas are due to Greenhouse and Mantel [19].

Suppose G is the distribution function for assay values in a healthy (control) population and F is the distribution function for diseased individuals (cases). Using as cutpoint for declaring 'diseased' the 95th percentile, $\xi_{0.95} = G^{-1}(0.95)$ would yield 95 per cent specificity for an assay result X deemed positive if $X > \xi_{0.95}$. The corresponding sensitivity, namely $P(\text{declare 'diseased'} \mid \text{diseased})$, would be $1 - F\{G^{-1}(0.95)\} = 1 - F(\xi_{0.95})$.

The first problem Greenhouse and Mantel considered was whether the assay X had promise. A good assay should have little overlap among cases and controls in order to be highly discriminating. Therefore Greenhouse and Mantel considered the null hypothesis $G^{-1}(0.95) \leq F^{-1}(0.10)$, which defined a good assay because the assay would have sensitivity at least

0.90 based on a cutpoint that assured specificity 0.95. The null hypothesis would be rejected in favour of $G^{-1}(0.95) > F^{-1}(0.10)$, an indication of a bad test with sensitivity less than 90%. Greenhouse and Mantel developed the necessary parametric theory and sample size requirements under the assumption that F and G were normally distributed. They also gave a non-parametric version of this test based on the distribution of percentiles in independent samples from G and F .

Greenhouse and Mantel then turned to the comparison of two diagnostic assays, X_1 and X_2 , with respective marginal distributions F_1 and F_2 in cases and G_1 and G_2 in controls. As before, we assume an assay is positive if it exceeds a cutpoint. Mantel and Greenhouse pointed out that a fair comparison of sensitivities of X_1 and X_2 could only be made if they were operating at the same specificity. The cutpoint (percentile) required to yield specificity 0.95, say, needed to be estimated from control samples and was thus itself a random variable. This additional variability needed to be taken into account when testing the null hypothesis of equal sensitivity:

$$1 - F_1(\xi_1) = 1 - F_2(\xi_2)$$

where ξ_1 and ξ_2 are the 95th percentiles of G_1 and G_2 , respectively.

Greenhouse and Mantel distinguished two types of sampling. The assays X_1 and X_2 could be studied in independent samples of cases and controls, or X_1 and X_2 could each be measured in pairs in the same samples of cases and controls. To cover the latter instance, let $F(x_1, x_2)$ be the joint distribution in cases with $F_1(x_1) = F(x_1, \infty)$ and $F_2(x_2) = F(\infty, x_2)$. Define the joint distribution G similarly for controls. The null hypothesis of equal sensitivity can be tested using $[\hat{F}_1\{\hat{G}_1^{-1}(0.95)\} - \hat{F}_2\{\hat{G}_2^{-1}(0.95)\}]\hat{V}^{-\frac{1}{2}}$ where \hat{V} estimates the variance and \hat{F}_1 , \hat{F}_2 , \hat{G}_1 , \hat{G}_2 are estimates of the marginal distributions. Greenhouse and Mantel gave results both for the case of normally distributed X_1 and X_2 and for the non-parametric analysis, and they treated both independent and paired assay sampling. For paired data, on n cases and m controls, we can express the variance V for empirical (non-parametric) distributions \hat{F}_1 , \hat{F}_2 , \hat{G}_1 , and \hat{G}_2 in terms of densities such as f_1 corresponding to F_1 as

$$\begin{aligned} & \frac{F_1(\xi_1)\{1 - F(\xi_1)\}}{n} + \frac{F_2(\xi_2)\{1 - F(\xi_2)\}}{n} - \frac{2F(\xi_1, \xi_2) - F_1(\xi_1)F_2(\xi_2)}{n} \\ & + \left(\frac{f_1(\xi_1)}{g_1(\xi_1)}\right)^2 \frac{(0.95)(0.05)}{m} + \left(\frac{f_2(\xi_2)}{g_2(\xi_2)}\right)^2 \frac{(0.95)(0.05)}{m} \\ & - \frac{2f_1(\xi_1)f_2(\xi_2)}{g_1(\xi_1)g_2(\xi_2)} \frac{F(\xi_1, \xi_2) - (0.05)^2}{m} \end{aligned}$$

In this expression the first three terms correspond to binomial variation assuming the percentiles ξ_1 and ξ_2 were known, but taking the sample correlations from pairing into account. The last three terms represent additive random variation resulting from the need to estimate the percentiles (cutpoints) ξ_1 and ξ_2 from data on non-diseased subjects. A related interpretation is to regard the first three terms in this expression as approximating the expected value of the conditional variance, given estimates of the cutpoints; the second three terms correspond to the variance of the conditional expectation given the empirical cutpoints.

Introduce labeled albumin
into compartment 1
and follow time course
of concentration levels

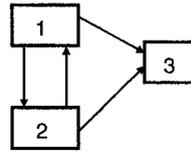


Figure 1. A three-compartment kinetic model for albumin.

This paper offered many helpful insights concerning methods for evaluating diagnostic tests and is widely cited and used by current researchers.

3. MODELS FOR THE INTERPRETATION OF EXPERIMENTS USING TRACER COMPOUNDS

The availability of radioactive labelling techniques in the 1950s enabled an explosion of biochemical and physiological research into the distribution and flows of proteins and other molecules throughout the body. For example, albumin labelled with I^{131} might be injected into the blood stream, represented as compartment 1 in Figure 1. In Figure 1, compartment 2 might represent the interstitial space, which is in communication with the vascular component, and compartment 3, which is an absorbing state, represents excretion of albumin in urine and feces. The diminution of the radioactive label from compartment 1 over time gave information on the size of the vascular compartment and compartment 2 and on the rate of flows among compartments. Mathematical models were central to the interpretation of such data, and physicists, such as Mones Berman at NIH, helped define the key assumptions underlying such models, develop deterministic solutions to model systems, and determine what kinds of experiments were needed and on which compartments in order to identify the model parameters [20].

Lewallen *et al.* [21] used such methods to study the effects of thyroid disease on albumin metabolism based on a 3-compartment model. Although this brilliant article represented the state of the art in 1959 and produced important physiologic findings, curve-fitting methods that account for measurement error were *ad hoc*: 'The slowest component was obtained first by a least squares fit of the terminal linear portion of the plot. The remainder of the curve was smoothed by eye, and the two additional components were obtained by serial subtraction in the usual way.'

In an important article that introduced biostatisticians to this field, Cornfield, Steinfeld (who became Surgeon General of the United States Public Health Service in 1969), and Greenhouse carefully reviewed the underlying assumptions of kinetic modelling in biological systems and then turned to the difficult issue of model fitting for the 3-compartment system in Figure 1 [22]. In particular they inquired why standard numerical approaches often failed.

The solution of deterministic differential equations for the amount of labelled substance in compartment 1 of Figure 1 at time t_i , $y(t_i)$, is $A_1 \exp(-\lambda_1 t_i) + A_2 \exp(-\lambda_2 t_i)$. Assuming that the measurements of $y(t_i)$ are subject to independent mean zero errors with variances $1/w_i$, Cornfield, Steinfeld, and Greenhouse sought to estimate A_1 , A_2 , λ_1 and λ_2 by minimizing the weighted squared deviations $S = \sum w_i \{y(t_i) - A_1 \exp(-\lambda_1 t_i) - A_2 \exp(-\lambda_2 t_i)\}^2$. It follows from symmetry considerations and an examination of the four estimating equations obtained by dif-

ferentiation with respect to the parameters that if $\hat{\theta} = (\hat{A}_1, \hat{A}_2, \hat{\lambda}_1, \hat{\lambda}_2)$ is a global minimum of S , then so is $\theta^* \equiv (A_1^*, A_2^*, \lambda_1^*, \lambda_2^*) = (\hat{A}_2, \hat{A}_1, \hat{\lambda}_2, \hat{\lambda}_1)$. Cornfield, Steinfeld, and Greenhouse credited Nathan Mantel for this observation, but they took the analysis further. If $\hat{\theta} = (\hat{A}_1, \hat{A}_2, \hat{\lambda}_1, \hat{\lambda}_2)$ and $\theta^* = (\hat{A}_2, \hat{A}_1, \hat{\lambda}_2, \hat{\lambda}_1)$, each minimize S , then $\partial S / \partial G = 0$ and the determinant $D = |\partial^2 S / \partial \theta^2| > 0$ at these two points. It follows that there must be at least one stationary point θ^{**} satisfying $\partial S / \partial \theta = 0$ between $\hat{\theta}$ and θ^* . Iterative methods with initial values of θ near θ^{**} will converge to a local minimum, a maximum, or a saddle point. Further analysis revealed that θ^{**} is likely to be a saddle point in most applications. At a saddle point, $D < 0$. From the continuity of D , it follows that there must be a point between $\hat{\theta}$ and θ^{**} and a point between θ^{**} and θ^* at which $D = 0$. Iterative procedures such as Newton's method fail when $D = 0$ because the matrix of derivatives of the estimating equations is not invertible there. The analysis thus revealed at least two points where standard iterative procedures fail and at least one stationary point θ^{**} which is typically a saddle point and to which standard iterative procedures with nearby starting values converge. The paper proceeds to suggest ways to avoid misleading solutions and failures of iterative methods.

For the practicing biostatistician, the paper lays out the basic principles of analysis for labeling experiments and highlights potential pitfalls of statistical curve-fitting.

4. MULTIVARIATE RELATIVE RISK FUNCTIONS FOR CASE-CONTROL STUDIES

In 1962, Cornfield considered how to estimate the risk of coronary heart disease (CHD) from cohort data as a function of blood pressure and cholesterol levels [23]. Assuming that $X_1 = \log_{10}$ (cholesterol) and $X_2 = \log_{10}$ (systolic blood pressure - 75) were bivariate normal with mean μ_1 and covariance Σ in those with CHD and with mean μ_0 and covariance Σ in subjects without CHD, he showed that the log odds of disease given X_1 and X_2 equaled $\alpha + \beta_1 X_1 + \beta_2 X_2$. Here

$$\alpha = \log_e [P(\text{CHD}) / \{1 - P(\text{CHD})\}] - \beta'(\mu_1 + \mu_2) / 2$$

and $\beta' = (\beta_1, \beta_2) = (\mu_1' - \mu_0') \Sigma^{-1}$. Cornfield pointed out that the logistic model above included quadratic terms and a product in X_1 and X_2 if the covariance Σ differed in those with and without disease. Note that the probability of disease in the cohort, $\pi_1 = P(\text{CHD})$, is needed to estimate α , but not the log relative odds parameters β_1 and β_2 . Cornfield substituted means $\hat{\mu}_1$ and $\hat{\mu}_0$ from CHD and non-CHD subjects, respectively, and a pooled estimate of covariance Σ from the two samples, into the previous formulas to estimate $\hat{\beta}_1 = 6.14$ and $\hat{\beta}_2 = \beta.29$. This analysis showed how combined elevations of blood pressure and cholesterol increased risk dramatically.

Subsequent workers pointed out that the logistic risk model for cohort data can arise from distributions other than the normal [24, 25]. This observation suggested that a more robust analysis would be based on the assumption that the log odds of disease was linear in covariates X in the general population, rather than the more stringent assumption of multivariate normality in diseased and non-diseased subjects. Walker and Duncan developed maximum likelihood procedures for cohort data under the logistic model [26], and Efron quantitated the loss of statistical efficiency incurred by using the logistic model when the multivariate normal discriminant model held [27].

Seigel and Greenhouse felt constrained by available contingency table methods for case-control studies to analyse effects of continuous exposures or multiple covariates, and they evaluated linear regression methods and logistic models based on Gaussian linear discrimination instead [28]. Indeed, they followed the lead of Cornfield [23] by assuming that the covariate distribution was normal in cases and (separately) in controls with respective means μ_1 and μ_0 and with common covariance Σ . They observed that true log relative odds, β_1 , could be estimated from case-control data with the formula $\hat{\beta}' = (\hat{\mu}'_1 - \hat{\mu}'_0)\Sigma^{-1}$ given by Cornfield; in this expression, $\hat{\mu}'_1$ and $\hat{\mu}'_0$ are estimated as the mean covariate values in cases and controls respectively, and Σ is estimated from the pooled within-group covariances. Siegel and Greenhouse also noted that the logistic intercept could not be estimated from case-control data because the probability of disease in the source population, π_1 , was usually unknown.

Siegel and Greenhouse were concerned that the normality assumptions might be misleading, and they analyzed the special case of a dichotomous covariate $X = 1$ or 0 . If the numbers of cases with $X = 1$ is a , the number with $X = 0$ is b , the number of controls with $X = 1$ is c , and the number of controls with $X = 0$ is d , then the maximum likelihood estimate of β is $\ln(ad/bc)$. Siegel and Greenhouse noted that the Gaussian discriminant model estimate of $\hat{\beta}$ in this case was $(ad - bc)T / \{ab(c + d) + cd(a + b)\}$, where $T = a + b + c + d$. Thus, the Gaussian discriminant model yields biased estimates of the odds ratio. The bias is not always severe, however. For example, with $a = 30$, $b = 20$, $c = 10$, and $d = 20$, the maximum likelihood estimate of β is $\log(30 \times 20 / 20 \times 10) = \log(3) = 1.099$, whereas the normal discriminant estimate is 1.1428. The former corresponds to an odds ratio of 3.00, whereas the latter corresponds to an odds ratio of $\exp(1.1428) = 3.136$.

At about the same time Siegel and Greenhouse were working on this topic, Anderson [29] considered the implications of assuming that the logistic risk model held in the general population, rather than assuming multivariate normality in cases and controls. For discrete covariates, X , he showed that the estimate $\hat{\beta}$ of log relative risk that maximized the retrospective likelihood from case-control data was the same estimate that one would obtain by pretending that the case-control sample represented a prospective cohort study. He also showed that the variance of $\hat{\beta}$ was the same as would be obtained in such a cohort study. Prentice and Pyke [30] later proved these results for continuous covariates. In 1973 Mantel showed that if one assumed that the logistic risk model held in the general population, then it would also hold in the selected case-control sub-sample of that population, but with an altered intercept [31].

Thus, there were two themes in the case-control literature just as in the cohort literature, one based on an assumption of multivariate normality in cases and controls [28], and another based on an assumption of a logistic risk model in the general population. The paper by Siegel and Greenhouse was important in identifying a need for regression methods for case-control data, showing how the normal discriminant approach could be used for this purpose, and warning the user about the impact of violations of assumptions in that model.

5. DISCUSSION

I have illustrated through a few examples Sam's ability to identify problems of practical importance and to develop the necessary theory and methods to address them. Much of his methodological work remains important today, because it addressed problems of lasting

practical relevance. In an obituary published in *AmStat News* [32], John Lachin and Joel Greenhouse wrote of Sam: 'However, if one asked Sam about the truly important work he was doing, he would inevitably talk about his scientific collaborations. For it was through the practice of statistics, he believed, that statisticians made their biggest impact on science, and it was through scientific collaborations that the important statistical problems were identified.'

ACKNOWLEDGEMENTS

I would like to thank Dr Clarice Weinberg for pointing out key references and quotations and the reviewers for helpful clarifications.

REFERENCES

1. Greenhouse SW. Some reflections on the beginnings and development of statistics in 'your father's NIH'. *Statistical Science* 1997; **12**:82–87.
2. Goldin A, Venditti JM, Humphreys SR, Dennis D, Mantel N, Greenhouse SW. An investigation into the synergistic antileukemic action of amethopterin and 6-mercaptopurine in mice. *Journal of the National Cancer Institute* 1955; **16**:129–138.
3. Goldin A, Venditti JM, Humphreys SR, Dennis D, Mantel N, Greenhouse SW. Studies on the toxicity and antileukemic action of 6-mercaptopurine in mice. *Annals of the New York Academy of Sciences* 1954; **60**: 251–266.
4. Goldin A, Mantel N, Greenhouse SW, Venditti JM, Humphreys SR. Effect of delayed administration of citrovorum factor on the antileukemic effectiveness of aminopterin in mice. *Cancer Research* 1954; **14**: 43–48.
5. Greenhouse SW, Geisser S. On methods in the analysis of profile data. *Psychometrika* 1959; **24**:95–112.
6. Halperin M, Greenhouse SW, Cornfield J, Zalokar J. Tables of percentage points for the Studentized maximum absolute deviate in normal samples. *Journal of the American Statistical Association* 1955; **50**:185–195.
7. Halperin M, Greenhouse SW. Note on multiple comparisons for adjusted means in the analysis of covariance. *Biometrika* 1958; **45**(Parts 1, 2):256–259.
8. Greenhouse SW. The role of hypothesis testing in clinical trials: the mental disorders. *Journal of Chronic Diseases* 1966; **19**:859–862.
9. Cornfield J, Halperin M, Greenhouse SW. An adaptive procedure for sequential clinical trials. *Journal of the American Statistical Association* 1969; **64**:759–770.
10. Dambrosia J, Greenhouse SW. Early stopping for sequential restricted tests of binomial distributions. *Biometrics* 1983; **39**:659–710.
11. Dambrosia J, Greenhouse SW. A diffusion process approximation approach to restricted sequential tests with early stopping. *Communications in Statistics: Sequential Analysis* 1984; **3**:213–230.
12. Seigel DC, Greenhouse SW. Validity in estimating relative risk in case-control studies. *Journal of Chronic Disease* 1973; **26**:219–225.
13. Mantel N, Greenhouse SW. What is the continuity correction? *The American Statistician* 1968; **22**(5):27–30.
14. Mantel N, Greenhouse SW. Equivalence of maximum likelihood and the method of moments in probit analysis. *Biometrics* 1965; **43**:154–157.
15. Gastwirth JL, Greenhouse SW. Estimating a common relative risk: application in equal employment. *Journal of the American Statistical Association* 1987; **82**:38–45.
16. Gastwirth JL, Greenhouse SW. Biostatistical concepts and methods in the legal setting. *Statistics in Medicine* 1995; **14**:1641–1653.
17. Dunn JE Jr, Greenhouse SW. Cancer diagnostic tests: principles and criteria for development and evaluation. Federal Security Agency, Public Health Service, #9, Government Printing Office 1950.
18. Dunn JE Jr, Greenhouse SW. The development and evaluation of cancer diagnostic tests. *Public Health Reports* 1953; **68**:880–884.
19. Greenhouse SW, Mantel N. The evaluation of diagnostic tests. *Biometrics* 1950; **6**:399–412.
20. Berman M, Schoenfeld R. Invariants in experimental data on linear kinetics and the formulation of models. *Journal of Applied Physics* 1956; **27**:1361–1370.
21. Lewallen CG, Berman M, Rall JE. Studies of iodoalbumin metabolism. I. A mathematical approach to kinetics. *Journal of Clinical Investigation* 1959; **386**:66–87.
22. Cornfield J, Steinfeld J, Greenhouse SW. Models for the interpretation of experiments using tracer compounds. *Biometrics* 1960; **16**:212–234.

23. Cornfield J. Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis. *Federation Proceedings* 1962; **21**(Part II. Suppl. II):58–61.
24. Cox DR. Some procedures connected with the logistic qualitative response curve. In *Festschrift for J. Neyman: Research Papers in Statistics*, David FM (ed.). Wiley: London, 1966; 55–71.
25. Day NE, Kerridge DF. A general maximum likelihood discriminant. *Biometrika* 1967; **23**:313–323.
26. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. *Biometrika* 1967; **54**:167–179.
27. Efron B. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* 1975; **70**:892–898.
28. Seigel DC, Greenhouse SW. Multiple relative risk functions in case-control studies. *American Journal of Epidemiology* 1973; **97**(5):324–331.
29. Anderson JA. Separate sample logistic discrimination. *Biometrika* 1972; **59**:19–35.
30. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979; **66**:403–411.
31. Mantel N. Synthetic retrospective studies and related topics. *Biometrics* 1973; **29**:479–486.
32. Lachin JM, Greenhouse JB. Obituary: Sam Greenhouse. *Amstat News* 2000; **282**:35–38.