# Deming and Survey Sampling

By

Wayne A. Fuller

Deming Lecture

JSM 2003

# W. Edwards Deming, Early Time Line

| | |
|---|---|
| 1900 | Born Sioux City, Iowa, October 14 |
| 1907 | Family moved to Wyoming |
| 1921 | U. of Wyoming, B.S. Electrical Engineering |
| 1924 | M.S. Colorado, Mathematics & Physics |
| 1928 | Ph.D. Yale, Mathematics & Physics |
| 1927-1939 | Fixed Nitrogen Research Laboratory, USDA |
| 1928-1942 | U.S.D.A. Graduate School |
| 1943 | *Statistical Adjustment of Data* (1938 mimeo) |
| 1939-1946 | Bureau of the Census |
| 1946-1993 | Consultant, New York U. |
| 1947 | First visit to Japan |
| 1950 | *Some Theory of Sampling* |

# Additional Dates

1950 Japan quality short course

1960 *Sample Design in Business Research*

1980 NBC documentary

1982 *Quality, Productivity and Competitive Position*

1986 *Out of Crisis*

1993 *The New Economics for Industry, Education and Government*

1993 Died December 20

# Discussion Topics

Least squares and raking

Analytic and enumerative surveys

# Least Squares

The application of least squares (*Phil. Mag.*1931)

On the application of least squares II (*Phil. Mag.*1934)

The Chi-test and curve fitting (*JASA* 1934)

On the application of least squares III (*Phil. Mag.*1935)

*Statistical Adjustment of Data*, 1943

W.G. Cochran.

for

[S]tatistical Laboratory

# Mistake of incorrectly assuming observations are error free

"Most texts do not warn the reader that problems may arise for which equations (14) [those that assume no errors in *x*] are not applicable; much less do they tell how to handle such a situation."

Deming (1931, p.154)

# Least Squares
# Errors - in - Variables

"Principal of least squares": Minimize

$$\sum_i w_i (X_i - x_i)^2$$

$$\hat{x}_i - \text{most probable value}$$

$$X_i - \text{observations}$$

$$w_i - \text{weight : inverse of variance}$$

# Least Squares: Constraints

The $x_i$ satisfy a law

Example: pressure, volume, temperature of a gas

Example: Three angles of a triangle

# Deming's Triangle Example

$X_i, \; i = 1, 2, 3$ Measured angles

$x_i, \; i = 1, 2, 3$ True angles

$$\sum_{i=1}^{3} x_i = 180^\circ$$

$$\hat{x}_i = X_i - w_i^{-1}\left(\sum_{j=1}^{3} X_j - 180^\circ\right)\left(\sum_{j=1}^{3} w_j^{-1}\right)^{-1}$$

# Curve Fitting

Constraints have parameters

$$\sum_{i=1}^{n} \left[ w_{xi}(X_i - x_i)^2 + w_{yi}(Y_i - y_i)^2 \right]$$

$$(X_i, Y_i) = (x_i, y_i) + (u_i, e_i)$$

$$(u_i, e_i) \sim \text{ind.}\left(\mathbf{0}, \text{diag}\left\{w_{xi}^{-1}, w_{yi}^{-1}\right\} s^2\right)$$

$$F(x_i, y_i; \mathbf{ß}) = 0, \qquad i = 1, 2, \ldots, n$$

# Curve Fitting, Example

Linear in ß constraint, $w_{xi} = w_{yi} = 1$

$$\text{Min} \sum_{i=1}^{n} \left[ \left( X_i - x_i \right)^2 + \left( Y_i - y_i \right)^2 \right]$$

subject to

$$y_i - b_o - b_1 x_i = 0, \quad i = 1, 2, \ldots, n$$

Nonlinear problem, $n + 2$ unknowns

# Curve Fitting (Nonlinear)

Initial estimates

Small errors; Taylor expansion

Inverse matrix

$$s^{-2}\sum_{i=1}^{n}\left[\left(X_i-\hat{x}_i\right)^2 w_{xi}+\left(Y_i-\hat{y}_i\right)^2 w_{yi}\right]\sim\ c_d^2$$

$$d=n-\dim \mathbf{ß}$$

# Regression and Least Squares

Haavelmo (*Econometrica* 1944)

Cowles Commission 1943-1947

- System of equations
- Error in equation
- Exogenous variables

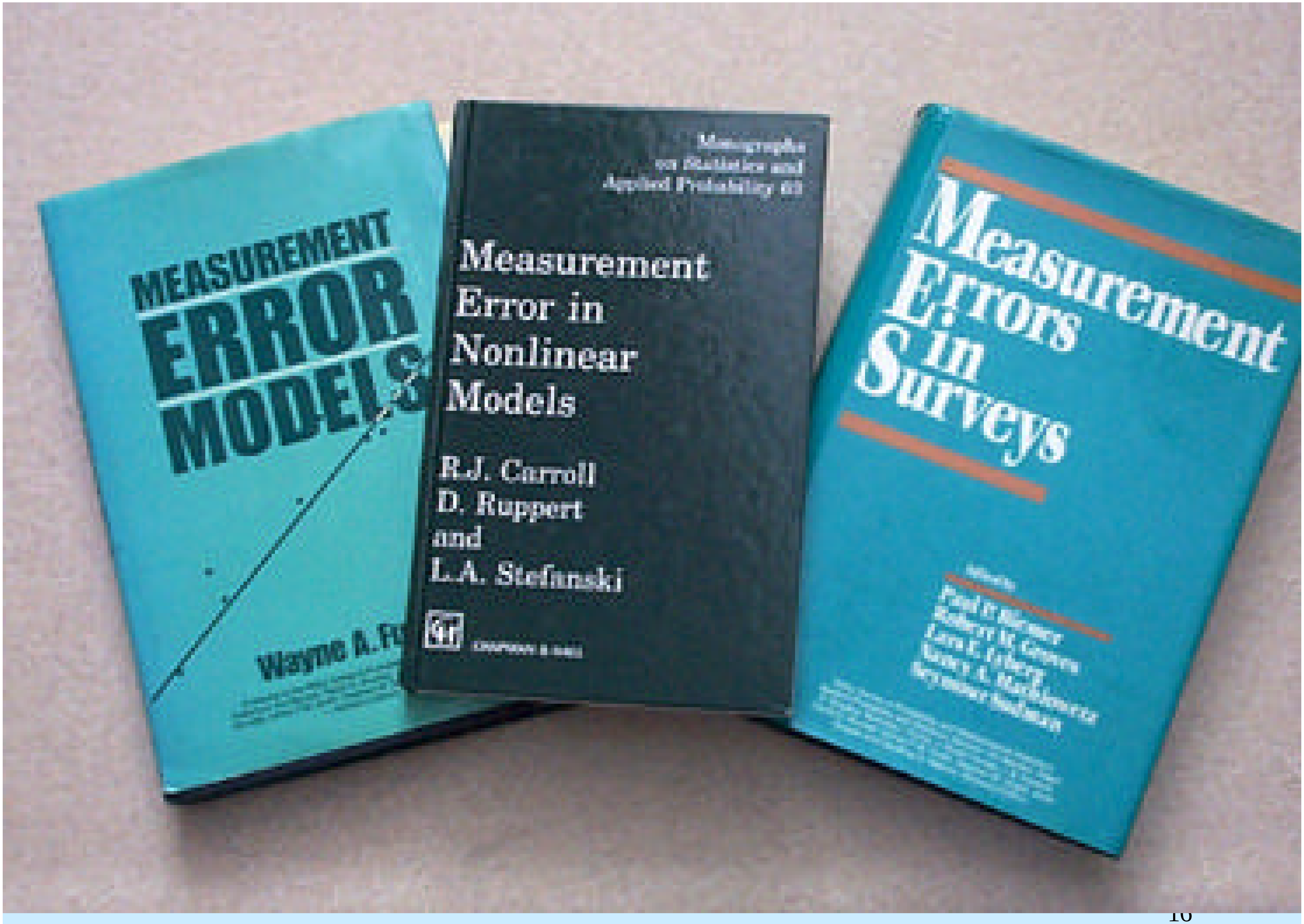
# Errors - in - Variables

$$F(x_i, y_i, \mathbf{ß}) = 0, \qquad i = 1, 2, \ldots, n$$

$$(X_i, Y_i) = (x_i, y_i) + (u_i, e_i)$$

$$(u_i, e_i) \sim ind(\mathbf{0}, \mathbf{S}_i)$$

$$\mathbf{S}_i \text{ known}$$


MEASUREMENT ERROR MODELS
Wayne A. Fu
Measurement Error in Nonlinear Models
R.J. Carroll
D. Ruppert
and
L.A. Stefanski
Measurement Errors in Surveys

# Some Activities for E-I-V Since 1943

Nonlinear, small error problem

Amemiya and Fuller (1988)

Cook and Stefanski (1994)

ODRPACK, NISS

# Activities Since 1943

Restrictions linear in $x$

Anderson (*AMS*, 1951), Maximum likelihood

Large $n$ and fixed (not small) covariance

Residual analysis

Random $x$

# Measurement Error Models

## Error in equation and measurement error

$$y_i = b_0 + x_i b_1 + a_i$$

$$(X_i, Y_i) = (x_i, y_i) + (u_i, v_i)$$

$$Y_i = \text{ß}_0 + x_i \text{ß}_1 + e_i, \quad e_i = a_i + v_i$$

$$(e_i, u_i) \sim ind.\left(\mathbf{0}, diag\left\{s_e^2, s_u^2\right\}\right) \; ind.\, of \;\; x_j$$

# Information for Identification

Instrumental variables, surrogates, indicators

Information on $s_u^2$, replication

Wu, L. (*JASA*, Dec 2002)

Ray Carroll, STATA

# Measurement Errors in Surveys

Measure magnitude

Minimize effect of error

Estimate variance of totals

# Measurement Error in Surveys

Mahalanobis (*Phil. Tr. Roy. Soc.* 1944, *JRSS* 1946)

Hansen, Hurwitz and Bershad (*ISI*, 1961)

Cochran (*Technometrics*, 1968)

Dalenius (*Int. Stat. Rev.* 1977)

*Measurement Errors in Surveys*, 1991

# Processing Errors

Deming and Geoffrey (*JASA*, 1941)

Deming, Tepping and Geoffrey (*JASA*, 1942)

Bureau of Census is a statistical factory

Control process for coding and punching

Five percent sample

Sampling:  little effect on quality

# Correlation Between Two Reports

| Variable | Correlation |
| --- | --- |
| Sex | 0.98 |
| Age 45-49 | 0.92 |
| Income | 0.85 |
| Unemployed | 0.77 |

# Adjustment of a Sampling Frequency Table

Deming and Stephan (*Ann. Math. Stat*. 1940)

Stephan (*Ann. Math. Stat*. 1942)

Sample table: $n_{ij},\ i = 1, \ldots, r,\ \ j = 1, \ldots, s$

Known (Census) margins

$$m_{i\cdot} = nN^{-1}N_{i\cdot}, \quad i = 1, 2, \ldots, r$$

$$m_{\cdot j} = nN^{-1}N_{\cdot j}, \quad j = 1, 2, \ldots, s$$

# Example: $2 \times 3$ table

| | | | | |
|---|---|---|---|---|
| $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{1\cdot}$ | $m_{1\cdot}$ |
| $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{2\cdot}$ | $m_{2\cdot}$ |
| $n_{\cdot 1}$ | $n_{\cdot 2}$ | $n_{\cdot 3}$ | $n_{\cdot\cdot}$ | |
| $m_{\cdot 1}$ | $m_{\cdot 2}$ | $m_{\cdot 3}$ | | $m_{\cdot\cdot}$ |

# Adjustment of a Table

Least squares: minimize

$$\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}^{-1}\left(n_{ij} - m_{ij}\right)^2$$

subject to

$$\sum_{j=1}^{s} m_{ij} = m_{i\cdot}, \quad i = 1, 2, \ldots, r$$

$$\sum_{i=1}^{r} m_{ij} = m_{\cdot j}, \quad j = 1, 2, \ldots, s$$

# Adjustment of a Table, I

Least squares: minimize

$$\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}^{-1}\left(n_{ij} - m_{ij}\right)^2$$

subject to

$$\sum_{j=1}^{s} m_{ij} = m_{i.}, \quad i = 1, 2, \ldots, r$$

$$\text{Solution}: \ \hat{m}_{ij} = n_{i.}^{-1} m_{i.} n_{ij}$$

# Adjustment of a Table, II

Least squares: minimize

$$\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}^{-1}\left(n_{ij} - m_{ij}\right)^2$$

subject to

$$\sum_{j=1}^{s} m_{ij} = m_{i.}, \qquad i = 1, 2, \ldots, r$$

$$\sum_{j=1}^{r} m_{ij} = m_{.j}, \qquad j = 1, 2, \ldots, s$$

# Iterative Proportional Fitting

Adjust to row totals

$$\hat{m}_{ij}^{(1)} = n_{i\cdot}^{-1} m_{i\cdot} n_{ij}, \quad i = 1, 2, \ldots, r$$

Adjust to column totals

$$\hat{m}_{ij}^{(2)} = n_{\cdot j}^{-1} m_{\cdot j} \hat{m}_{ij}^{(1)}, \quad j = 1, 2, \ldots, s$$

Adjust to row totals

$$\hat{m}_{ij}^{(3)} = n_{i\cdot}^{-1} m_{i\cdot} \hat{m}_{ij}^{(2)}, \quad i = 1, 2, \ldots, r$$

Continue to convergence

# Iterative Proportional Fitting

Ireland and Kullback (*Biometrika*, 1968)

Bishop, Fienberg and Holland (1975)

# Raking

Remains a very important procedure

- Few iterations

Regression related procedures

- Deville and Särndal (*JASA*, 1992)

# Some Sampling Landmarks

| | |
|---|---|
| 1926 | ISI Commission |
| 1934 | Neyman (*JRSS*) On two different aspects of the representative method: the method of stratified selection and the method of purposive selection. |
| 1937 | U.S. Census sample of unemployment |
| 1937 | Neyman lectures at USDA Graduate School |
| 1940 | Sampling in Census of Population. Stephan, Deming and Hansen (*JASA* 1940) |
| 1947 | Sample for Greece. Jessen, Blythe, Kempthorne and Deming (*JASA*, 1947) |

# Sampling Books

1949 Yates. *Sampling Methods for Censuses and Surveys*

1950 Deming. *Some Theory of Sampling*

1953 Hansen, Hurwitz and Madow. *Sample Survey Methods and Theory* Volume I and Volume II

1953 Cochran. *Sampling Techniques*

1953 P.V. Sukhatme. *Sampling Theory of Surveys with Applications*

# Deming’s Chapter One

In general, *statistical work consists first of all of determining what kinds of statistical information would be useful for the ends in view; of deciding whether the desired information can be obtained at all or at reasonable cost; and then of procuring this information at the lowest possible cost, and interpreting it in a form that assists rational decisions and adds to knowledge.*

# Deming’s Steps in Taking a Survey

Define information and universe

Create collection instrument, instructions

Lay out survey design

Carry out survey

Interpret and publish results

# Analytic and Enumerative

Deming and Stephan (*JASA*, 1941) On Interpretation of Censuses as Samples

Deming (*JASA*, 1953) On the Distinction Between Enumerative and Analytic Surveys

Deming (1950) *Some Theory of Sampling* (Ch.7)

# JASA (1941)

Quotes from Shewhart, Truesdell, Hotelling, Hauser

Current population is a function of chance variations in births, migrations, etc.

“population . . . is but one of the infinity of populations that will result by chance from the same underlying social and economic cause systems.”

# Census

Census is measure for one population

Generalizations must recognize fluctuations

# Statistical Implications

Small cells: large variance for analytic

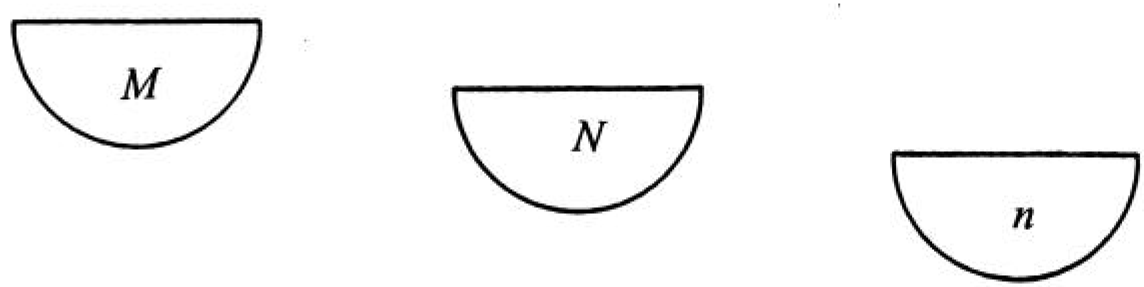
Sampling satisfactory: samples lead to same action as a complete count

Sampling can be superior: speed

# JASA (1953)

"*Purpose of paper*. Statistical data are supposedly collected to provide a rational basis for action. The action may call for the enumerative interpretation of the data, or it may call for the analytic interpretation."

# Deming’s Illustration




| SUPPLY | LOT-CONTAINER | SAMPLE-CONTAINER |
| --- | --- | --- |
| Initially contains | Into which are placed | Into which are placed |
| $Mp$ red | $x$ red | $r$ red |
| $Mq$ white | $N-x$ white | $n-r$ white |
| $M$ total | $N$ total | $n$ total |

# Assumptions

Lot chosen at random from process

M very large (for our purposes)

Random sample from lot

# Analytic

Purpose: To estimate the proportion in the *supply* (*cause system*)

$$\hat{p} = n^{-1} r$$

$$E\{\hat{p}\} = p$$

$$V\{\hat{p} - p\} = n^{-1} p(1 - p)$$

$r =$ number of red

# Enumerative

Purpose: To estimate the proportion in the *lot-container* ($\mathcal{F}$)

$$\hat{P} = n^{-1} r$$

$$E\{\hat{P} \mid \mathcal{F}\} = P = N^{-1} x$$

$$V\{(\hat{P} - P) \mid \mathcal{F}\} = (1 - N^{-1} n) n^{-1} P (1 - P)$$

# Enumerative

Number of individuals in a congressional district

Acres of corn harvested in US in 2002

# Analytic

Effect of prenatal care on maternal death-rate

Effect of income on consumption of meat

# Complex Analysis of Complex Surveys

- Subject matter model
  - Variables, relationship, error structure
  - Nature of generalizations
- Effect of survey design
  - Weights
  - Stratum, cluster
- Estimation and Design

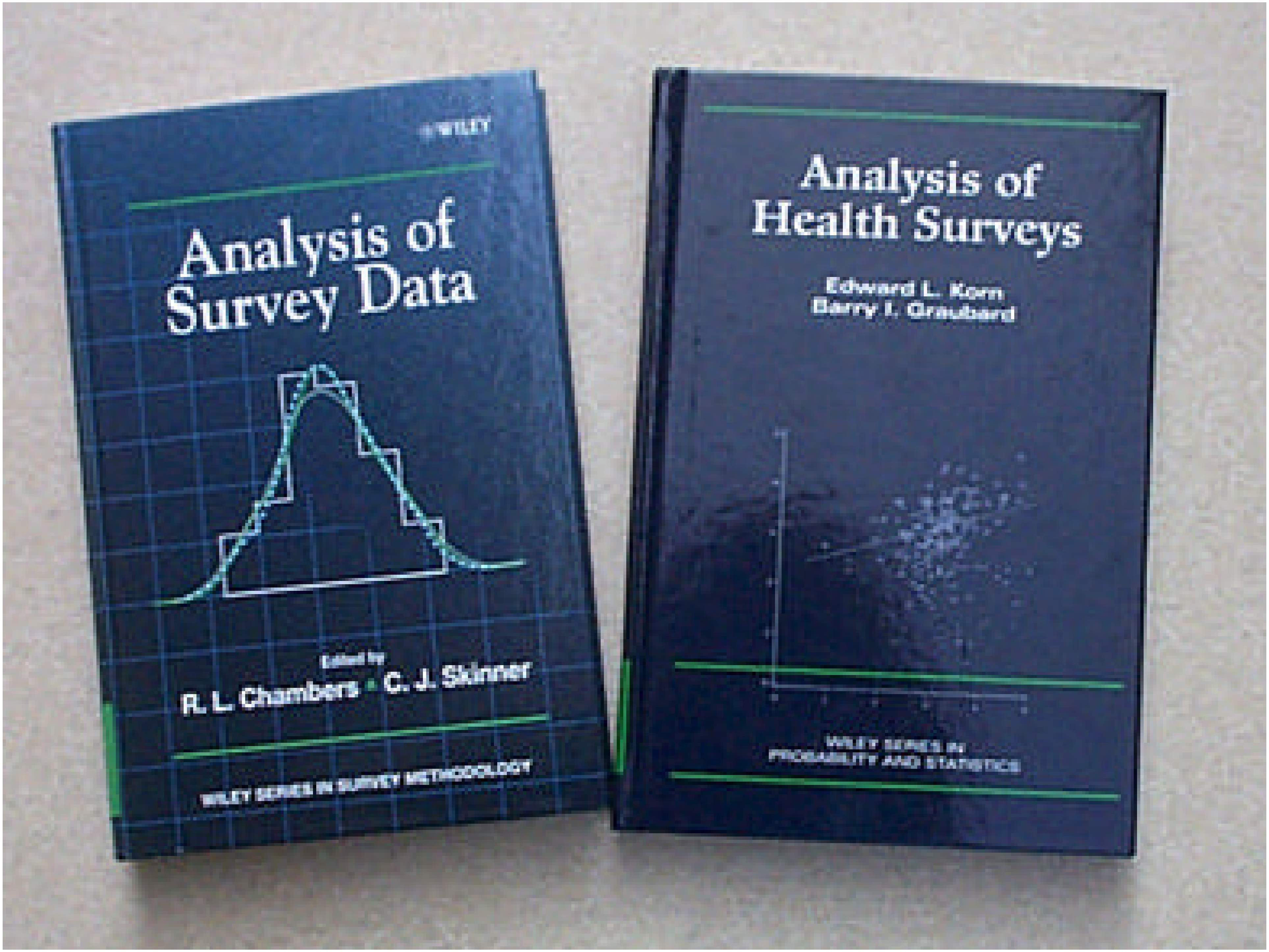
# Analytic Uses of Survey Data

Skinner, Holt and Smith (Eds.) (1989) *Analysis of Complex Surveys*

Thompson (1997) *Theory of Sample Surveys*

Korn and Graubard (1999) *Analysis of Health Surveys*

Chambers and Skinner (Eds.) (2003) *Analysis of Survey Data*


WILEY
Analysis of Survey Data
Edited by
R. L. Chambers • C. J. Skinner
WILEY SERIES IN SURVEY METHODOLOGY
Analysis of Health Surveys
Edward L. Korn
Barry I. Graubard
WILEY SERIES IN PROBABILITY AND STATISTICS

# Conclusion

Provide information as *basis for action*

Use resources wisely to meet objective

Communicate results