Title: Statistics, Uncertainty, and Transmitted Variation

Author(s): Wendelberger, Joanne Roth

Intended for: invited seminar, U. of Notre Dame

Issued: 2014-11-05

Disclaimer:

Los Alamos National Laboratory, an affirmative action/equal opportunity employer,is operated by the Los Alamos National Security, LLC for the National NuclearSecurity Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Departmentof Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Statistics, Uncertainty, and Transmitted Variation

Joanne R. Wendelberger, Ph.D.

Group Leader, Statistical Sciences Group

Los Alamos National Laboratory

joanne@lanl.gov

Technical Seminar, U. of Notre Dame, Notre Dame, IN, 2014

UNCLASSIFIED

Abstract

The field of Statistics provides methods for modeling and understanding data and making decisions in the presence of uncertainty. When examining response functions, variation present in the input variables will be transmitted via the response function to the output variables. This phenomenon can potentially have significant impacts on the uncertainty associated with results from subsequent analysis. This presentation will examine the concept of transmitted variation, its impact on designed experiments, and a method for identifying and estimating sources of transmitted variation in certain settings.

Introduction

- **Statistics** is sometimes referred to as the Science of **Uncertainty**.
- In recent years, increasing interest in uncertainty has led to the interdisciplinary field of Uncertainty Quantification (UQ) which focuses on understanding uncertainty throughout the modeling process.
- Modern Statistics and UQ draw on many statistical ideas that have evolved over the past century.

Uncertainty Quantification (UQ)

UQ is "the process of quantifying uncertainties associated with model calculations of true, physical quantities of interest, with the goals of accounting for all sources of uncertainty and quantifying the contributions of specific sources to the overall uncertainty."

National Research Council (2012):

“All models are wrong, but some are useful.”

Box, G.E.P. (1979), “Robustness in the Strategy of Scientific Model Building, in *Robustness in Statistics*, ed. By R. L. Launer and G. N. Wilkinson.

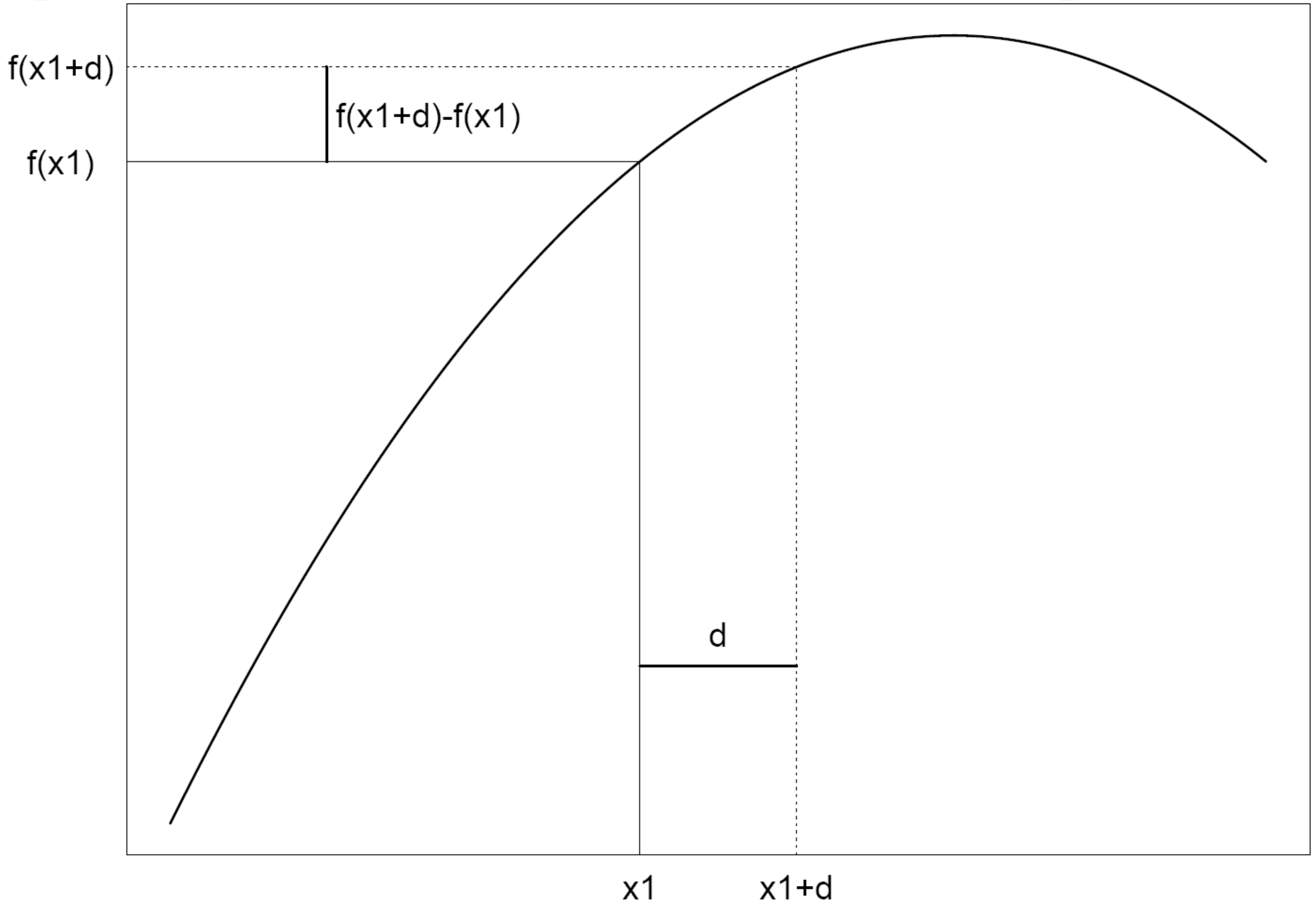
Ideas explored by Box and Wendelberger on Uncertainty and Transmitted Variation

- This work originally appeared in a U. of Wisconsin-Madison Ph.D. thesis supervised by George Box (Wendelberger, 1991) and two articles in the JSM Proceedings (1992, 1993).
- "Uncertainty in Designed Experiments" appeared in *Quality Engineering (2010)* in honor of George's 90th Birthday.
- "Variation in Controlled Experimental Variables" is to appear in a special issue of *Quality Technology & Quantitative Management,* A Tribute to George Box, (2015).

UNCLASSIFIED

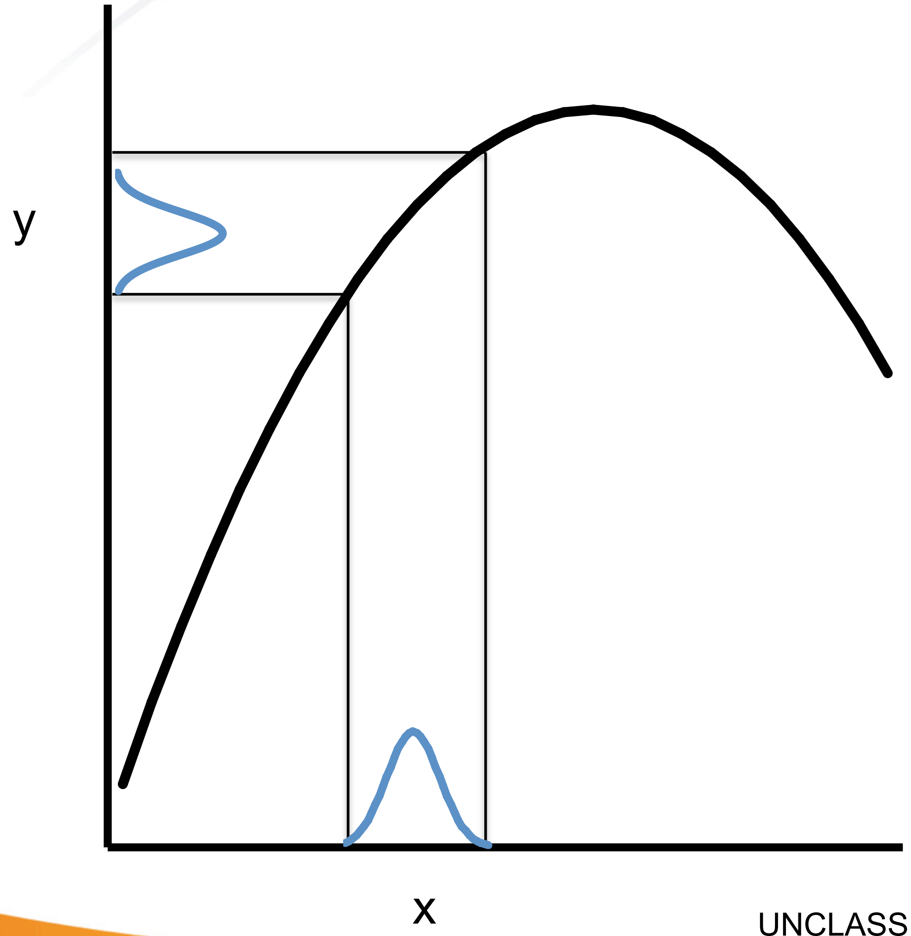
,

Transmitted Variation



UNCLASSIFIED

Transmitted Variation



- Uncertainty in an input variable *x* is transmitted to an output variable *y* via an underlying response function.

UNCLASSIFIED

Propagation of Error

Suppose $x_1, \ldots, x_p$ have independent errors

with variances $\sigma_1^2, \ldots, \sigma_p^2,$

$$y = f(x_1, \ldots, x_p) + \epsilon,$$

and

$$Var(\epsilon) = \sigma_0^2.$$

Then, a first order variance approximation is given by

$$V_{\tilde{x}}(y) = f_1^2\sigma_1^2 + \ldots + f_p^2\sigma_p^2 + \sigma_0^2,$$

where

$$f_i = \left.\frac{\partial f}{\partial x_i}\right|_{\tilde{x}}.$$

See Deming (1943), Ku (1969).

UNCLASSIFIED

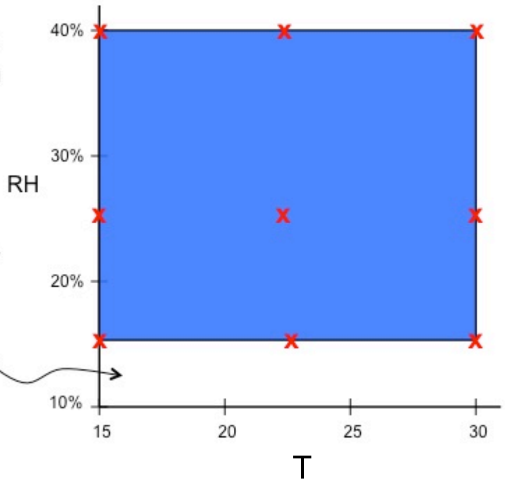
Uncertainty in Designed Experiments

- What happens to designed experiments in the presence of variation in the experimental inputs?
- Variation in controlled experimental variables induces variation in the measured responses.
- This can impact properties of the experimental design such as orthogonality and efficiency of estimates.

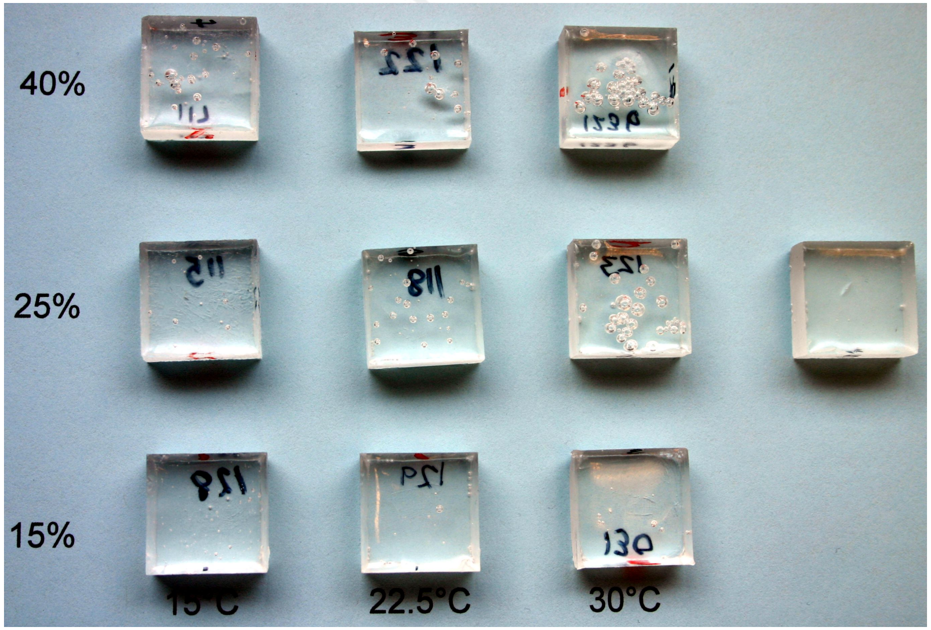
Polymer experiment

- 6 replicates at each condition

Design Region



Environmental chamber could not control RH reliably below 15%. Hardness and density samples also being prepared at 0% RH

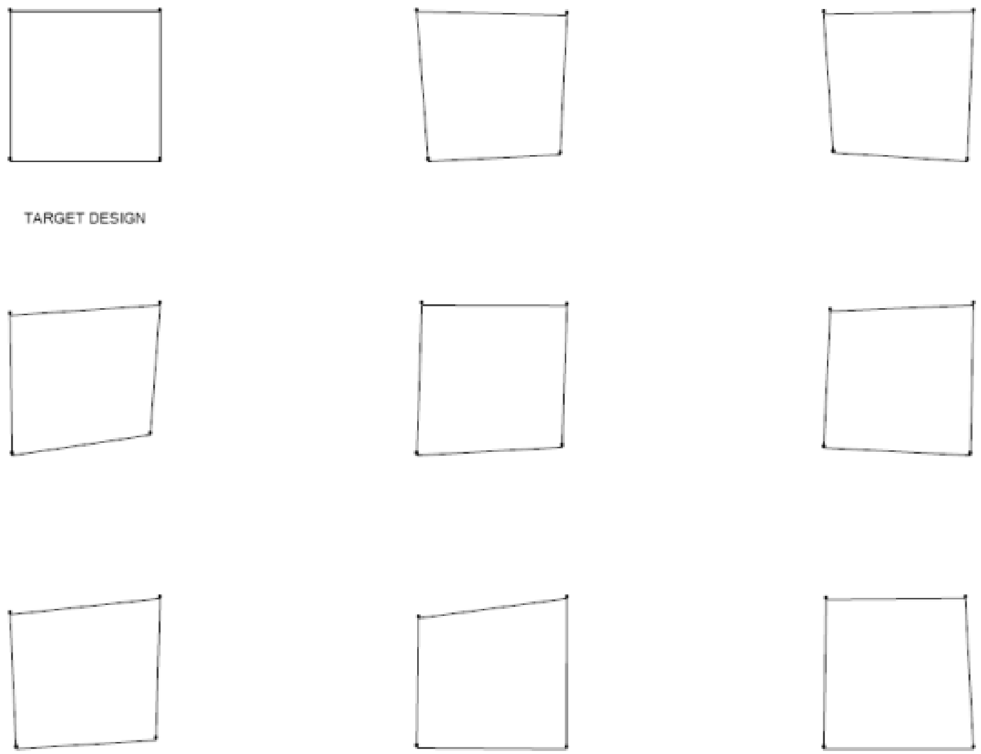


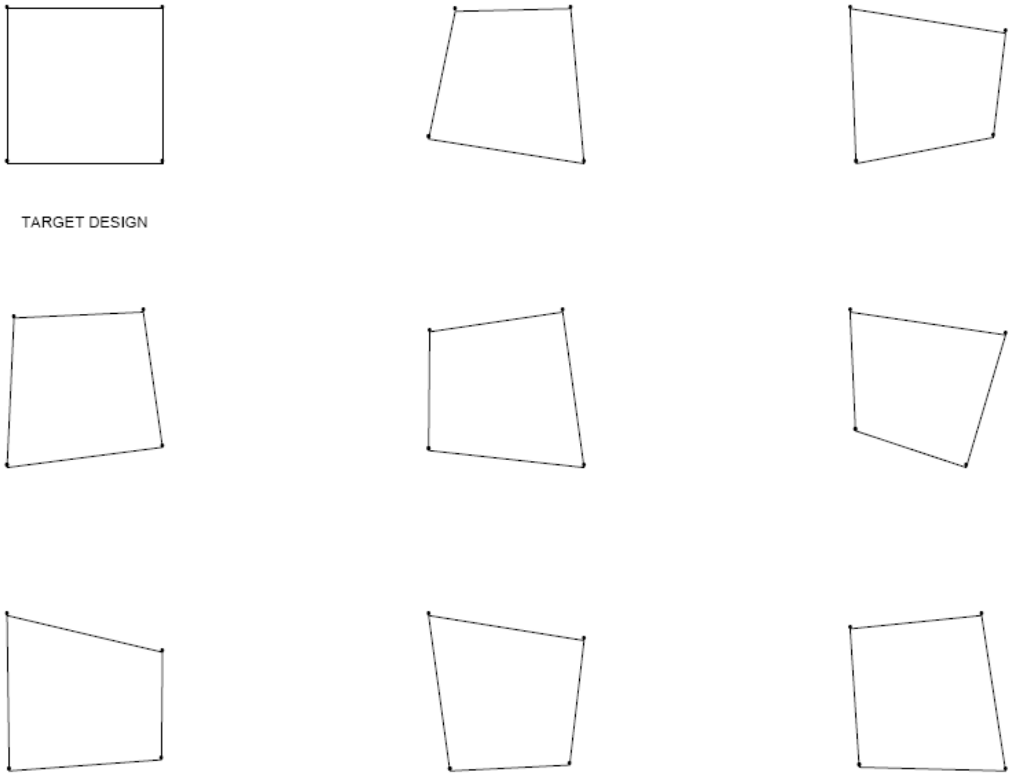
Polymer samples prepared at varying temperatures and relative humidity values.

Coatings Experiment

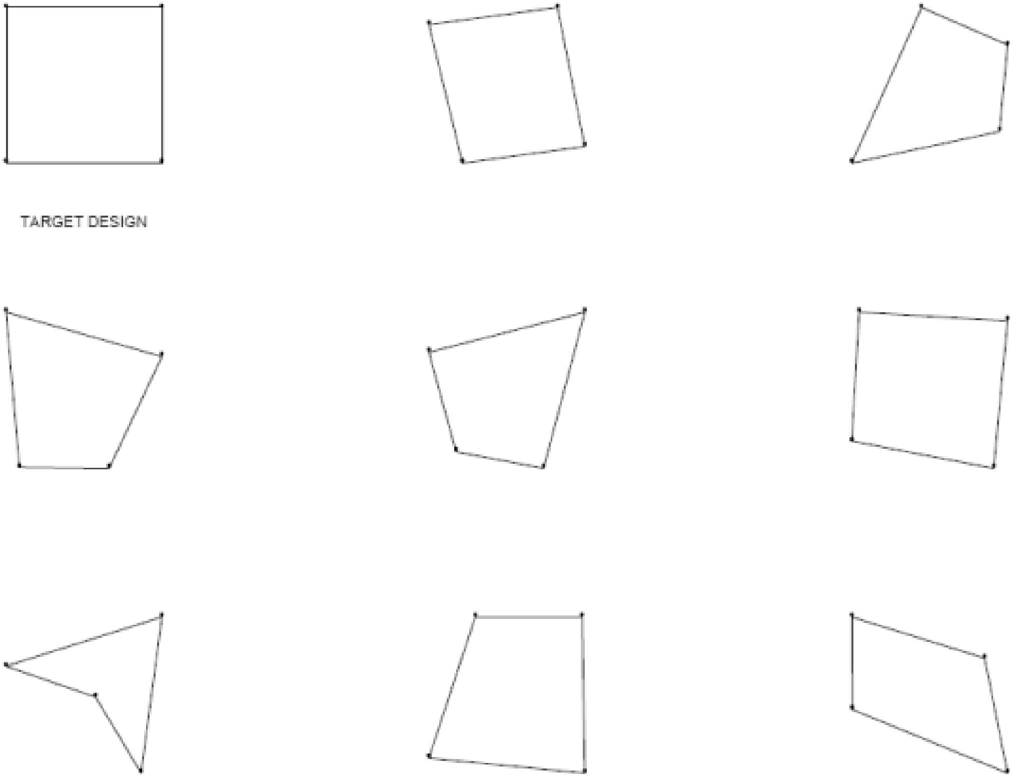
- Test panels coated under different experimental conditions
- Several different experimental factors
- Multiple responses associated with the quality of the coating
- Temperature difficult to control
- Humidity difficult to control

UNCLASSIFIED


TARGET DESIGN



UNCLASSIFIED


TARGET DESIGN

Variation in Controlled Experimental Variables

- Deviations in the factor settings lead to changes in the experimental region.
- Variation in the factor settings will be transmitted to the response variable.
- With sufficient data, unknown deviations can be identified and estimated for response functions with nonzero second derivative.

Early Work on Linear Models, Subject to Input Error:

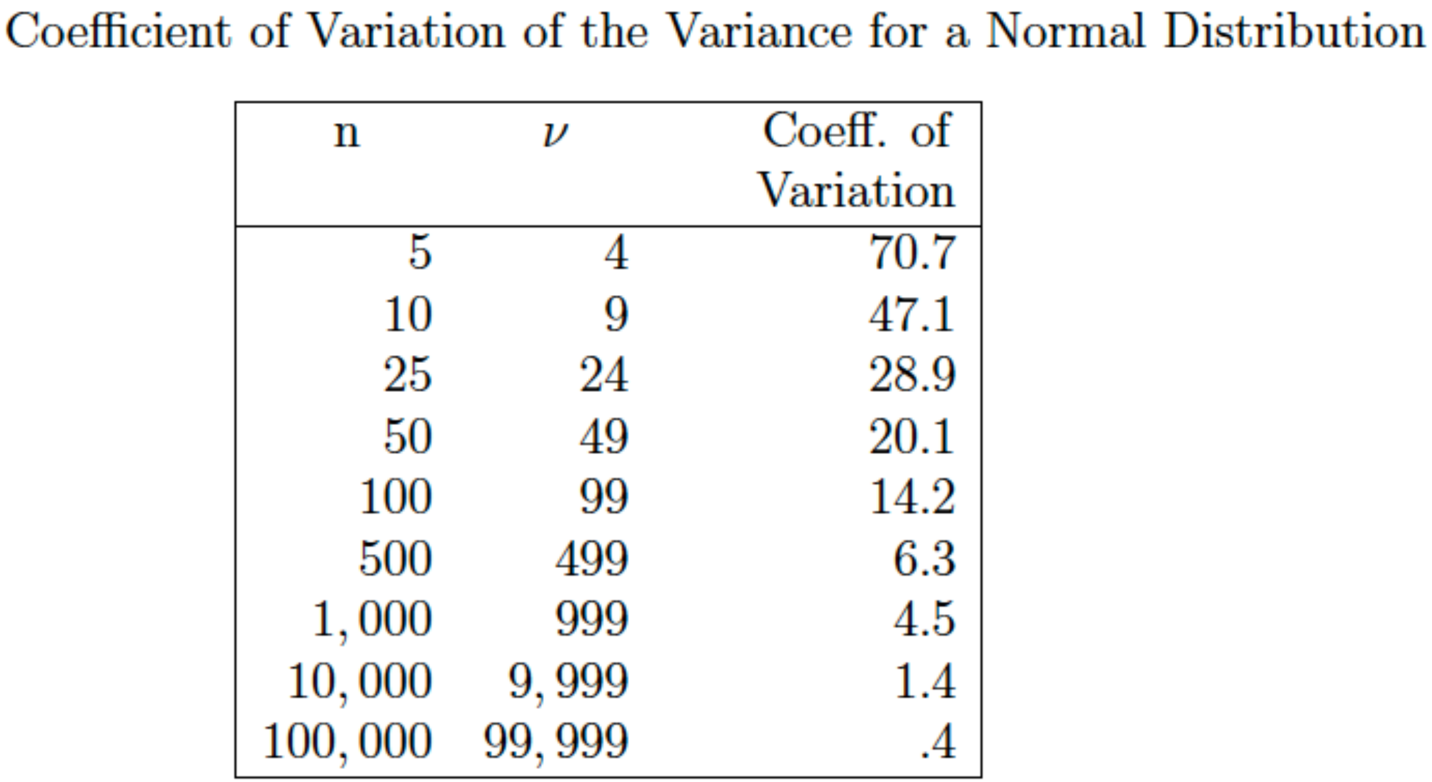
- Berkson (1950)
- Madansky (1959)
- Box (1963)
- Extensive errors in variables literature

Dispersion and Modeling of Variances

- Taguchi (1985), inner and outer arrays, S/N ratios
- Leon et al (1987), PERMIAs
- Box and Fung (1986), recognized relationship between response variance and gradient vector

UNCLASSIFIED

Variances are Variable!



Coefficient of Variation of the Variance for a Normal Distribution

n	ν	Coeff. of Variation
5	4	70.7
10	9	47.1
25	24	28.9
50	49	20.1
100	99	14.2
500	499	6.3
1,000	999	4.5
10,000	9,999	1.4
100,000	99,999	.4

Need 801 samples to obtain a Coefficient of Variation of 5%.

Transformations

- Bartlett and Kendall (1946) examined the use of a log transformation to stabilize variances.
- Box and Cox (1964) proposed a class of power transformations that are widely used in diverse application areas.

Three Cases

Consider three cases:

1. On-target experiment

2. Experiment with known deviations

3. Experiment with unknown deviations

Transmitted Variation

- Can compare relative efficiencies of on-target, known deviation, and unknown deviation cases.
- Can use variation in the response to identify sources of transmitted variation if underlying function has sufficient curvature.
- Can use maximum likelihood to obtain estimates of location and dispersion effects, which can then be used to estimate sources of transmitted variation.

Impact

Suppose ϵ has mean 0 and variance σ^2, and D is a deviation matrix where entries in column i are independent with mean 0 and variance σ_i^2.

Case 1: On-Target

$$y = X\beta + \epsilon.$$

$$b = (X'X)^{-1}X'y.$$

$$E(b) = \beta.$$

$$V(b) = (X'X)^{-1}\sigma^2.$$

Case 2: Unknown Deviations

$$y = (X + D)\beta + \epsilon,$$

$$y = X\beta + \alpha,$$

$$\text{where } \alpha = D\beta + \epsilon,$$

$$b = (X'X)^{-1}X'y.$$

$$E(b) = \beta.$$

$$V(b) = (X'X)^{-1}\sigma_\alpha^2,$$

$$\text{where } \sigma_\alpha^2 = \sigma^2 + \textstyle\sum_{i=1}^{p} \beta_i^2\sigma_i^2.$$

UNCLASSIFIED

Impact

Now suppose that the actual values $Z = X + D$ are known.

Case 2: Unknown Deviations

$$y = (X + D)\beta + \epsilon,$$

$$y = X\beta + \alpha,$$

$$\text{where } \alpha = D\beta + \epsilon$$

$$b = (X'X)^{-1}X'y.$$

$$E(b) = \beta.$$

$$V(b) = (X'X)^{-1}\sigma_{\alpha}^2,$$

$$\text{where } \sigma_{\alpha}^2 = \sigma^2 + \textstyle\sum_{i=1}^{p} \beta_i^2\sigma_i^2.$$

Case 3: Known Deviations

$$y = (X + D)\beta + \epsilon,$$

$$y = Z\beta + \epsilon.$$

$$b = (Z'Z)^{-1}Z'y.$$

$$E(b) = \beta.$$

$$V(b|Z) = (Z'Z)^{-1}\sigma^2.$$

$$V(b) = E_Z[(Z'Z)^{-1}]\sigma^2.$$

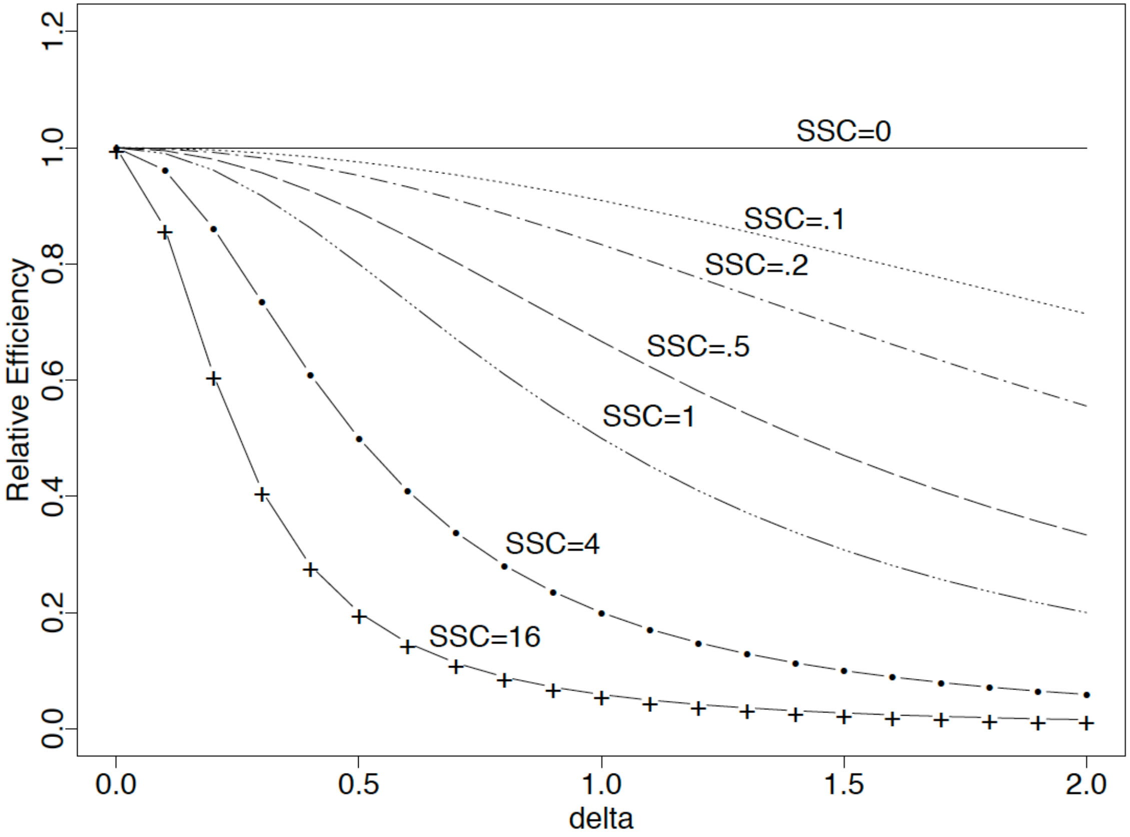
Relative Efficiencies

$$\text{R.E. of 2 to 1} = \frac{Var(b_j) \text{ for Case 1}}{Var(b_j) \text{ for Case 2}} = \frac{\sigma^2}{\sigma^2 + \sum_{i=1}^{p} \beta_i^2 \sigma_i^2}.$$

$$\text{R.E. of 3 to 1} = \frac{Var(b_j) \text{ for Case 1}}{Var(b_j|Z) \text{ for Case 3}} = \frac{[(X'X)^{-1}]_{jj}}{[(Z'Z)^{-1}]_{jj}}.$$

$$\text{R.E. of 3 to 2} = \frac{Var(b_j) \text{ for Case 2}}{Var(b_j|Z) \text{ for Case 3}} = \frac{[(X'X)^{-1}]_{jj}(\sigma^2 + \sum_{i=1}^{p} \beta_i^2 \sigma_i^2)}{[(Z'Z)^{-1}]_{jj}\sigma^2}.$$

Impact of Unknown Deviations



Effect of having deviations with unknown values. Each curve represents a particular sum of squared coefficients (SSC=$\sum_{i=1}^{p} \beta_i^2$) and shows how the relative efficiency of an experiment with unknown deviations to the on-target case changes for varying size $\delta = \sigma_x/\sigma$, assuming equal deviation variance σ_x.

UNCLASSIFIED

Identification

Use the following error transmission model, with response function (or quadratic approximation) *g*, and an additive error term *E*.

$$\ln s_u^2 = \ln(\sigma_0^2 + \sum_{i=1}^{p} g_{iu}^2 \sigma_i^2) + E_u,$$

which can be estimated using an iterative nonlinear least squares algorithm to obtain estimates of the error variances associated with each input.
i.e., use propagation of error in reverse.

Preliminary estimates of the parameters may be obtained using the untransformed model

$$s_u^2 = \sigma_0^2 + \sum_{i=1}^{p} g_{iu}^2 \sigma_i^2 + E_u$$

and simple linear regression.

Resulting estimates and approximate standard errors are used to tentatively identify the model.

UNCLASSIFIED

Estimation

- After tentatively identifying the model, can employ maximum likelihood techniques (or other methods) to obtain final estimates.
- Estimate the location and dispersion parameters.
- Then use the location and dispersion estimates and the transmitted variation model to estimate the sources of transmitted variation.
- Can compute standard errors from the asymptotic likelihood theory, or other techniques.

Dempster (1980), Efron (1978)

Estimation Example

In an experiment with n settings of the experimental variable x_u, $(u = 1, \ldots, n)$ each repeated m times $(i = 1, \ldots, m)$, we obtain response data y_{ui}, $(u = 1, \ldots, n)$, $(i = 1, \ldots, m)$. The probability density function of the response data, assuming a normal distribution on the error term e_{ui} is $p(y_{11}, y_{12}, \ldots, y_{1m}, \ldots, y_{n1}, \ldots, y_{nm})$

$$= \prod_{u=1}^{n} \left\{ \left(\frac{1}{\sqrt{2\pi}\gamma_u} \right)^m \exp\left(- \frac{\sum_{i=1}^{m} (y_{ui} - \beta_0 - \beta_1 x_u - \beta_{11} x_u^2)^2}{2\gamma_u^2} \right) \right. \qquad [26]$$

Let $L(\theta|y)$ denote the likelihood of the parameter vector $\theta = (\beta_0, \beta_1, \beta_{11}, \gamma_u,\ u = 1, \ldots, n)$ given the data $y_{11}, \ldots, y_{1m}, \ldots, y_{n1}, \ldots, y_{nm}$. Then the log likelihood of the parameters $\beta_0, \beta_1, \beta_{11},\ \gamma_u,\ u = 1, \ldots, n$ is

$$l(\theta|y) = \ln L(\theta|y) = \ln p(y|\theta)$$

$$= \sum_{u=1}^{n} \left\{ -\frac{m}{2} \ln(2\pi) - \frac{m}{2} \ln \gamma_u^2 - \frac{1}{2} \frac{\sum_{i=1}^{m} (y_{ui} - \beta_0 - \beta_1 x_u - \beta_{11} x_u^2)^2}{\gamma_u^2} \right\}. \qquad [27]$$

The first term inside the braces is constant and can be ignored. The third term can be reexpressed by subtracting off and adding $\bar{y}_u$ to each term in the sum. Thus,

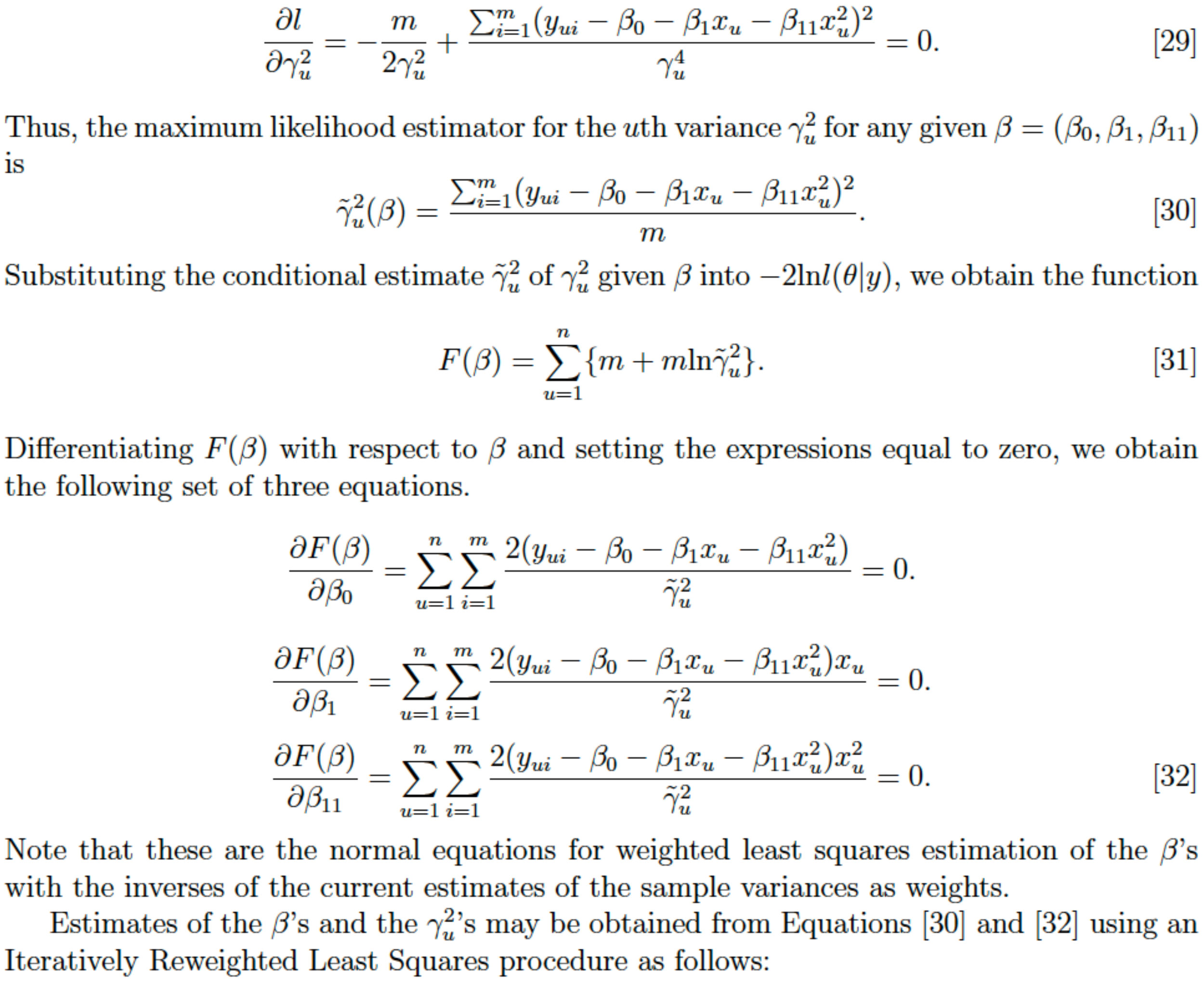
$$-2 \ln l(\theta|y) = m \sum_{u=1}^{n} \ln \gamma_u^2 + \sum_{u=1}^{n} \frac{\sum_{i=1}^{m} (y_{ui} - \bar{y}_u)^2 + m(\bar{y}_u - \beta_0 - \beta_1 x_u - \beta_{11} x_u^2)^2}{\gamma_u^2}$$

$$= m \sum_{u=1}^{n} \ln \gamma_u^2 + (m-1) \sum_{u=1}^{n} \frac{s_u^2}{\gamma_u^2} + m \sum_{u=1}^{n} \frac{(\bar{y}_u - \beta_0 - \beta_1 x_u - \beta_{11} x_u^2)^2}{\gamma_u^2}.$$

[28]

The likelihood breaks up into three sums, with the sample variances and the sample means appearing in separate terms.

Estimation Example



$$\frac{\partial l}{\partial \gamma_u^2} = -\frac{m}{2\gamma_u^2} + \frac{\sum_{i=1}^m (y_{ui} - \beta_0 - \beta_1 x_u - \beta_{11} x_u^2)^2}{\gamma_u^4} = 0. \qquad [29]$$

Thus, the maximum likelihood estimator for the uth variance γ_u^2 for any given $\beta = (\beta_0, \beta_1, \beta_{11})$ is

$$\tilde{\gamma}_u^2(\beta) = \frac{\sum_{i=1}^m (y_{ui} - \beta_0 - \beta_1 x_u - \beta_{11} x_u^2)^2}{m}. \qquad [30]$$

Substituting the conditional estimate $\tilde{\gamma}_u^2$ of γ_u^2 given β into $-2\ln l(\theta|y)$, we obtain the function

$$F(\beta) = \sum_{u=1}^n \{m + m\ln\tilde{\gamma}_u^2\}. \qquad [31]$$

Differentiating $F(\beta)$ with respect to β and setting the expressions equal to zero, we obtain the following set of three equations.

$$\frac{\partial F(\beta)}{\partial \beta_0} = \sum_{u=1}^n \sum_{i=1}^m \frac{2(y_{ui} - \beta_0 - \beta_1 x_u - \beta_{11} x_u^2)}{\tilde{\gamma}_u^2} = 0.$$

$$\frac{\partial F(\beta)}{\partial \beta_1} = \sum_{u=1}^n \sum_{i=1}^m \frac{2(y_{ui} - \beta_0 - \beta_1 x_u - \beta_{11} x_u^2)x_u}{\tilde{\gamma}_u^2} = 0.$$

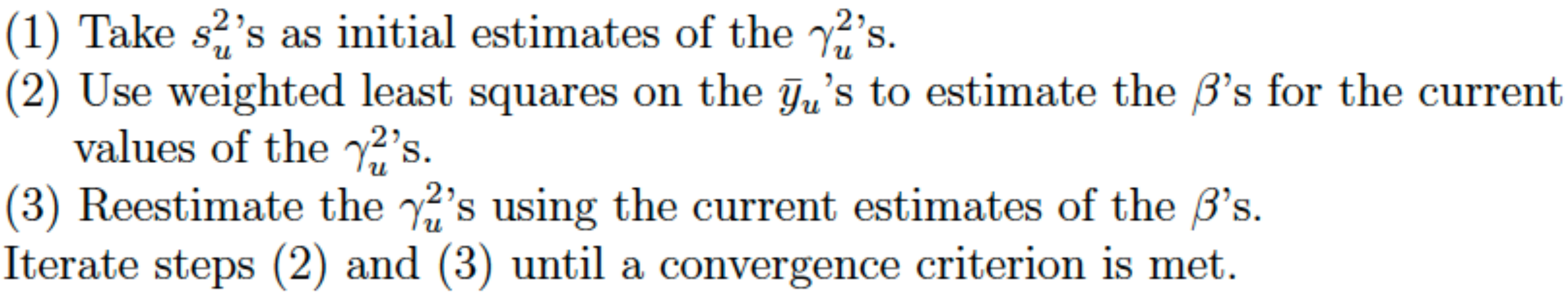
$$\frac{\partial F(\beta)}{\partial \beta_{11}} = \sum_{u=1}^n \sum_{i=1}^m \frac{2(y_{ui} - \beta_0 - \beta_1 x_u - \beta_{11} x_u^2)x_u^2}{\tilde{\gamma}_u^2} = 0. \qquad [32]$$

Note that these are the normal equations for weighted least squares estimation of the β's with the inverses of the current estimates of the sample variances as weights.

Estimates of the β's and the γ_u^2's may be obtained from Equations [30] and [32] using an Iteratively Reweighted Least Squares procedure as follows:

Estimation Example

Iteratively Reweighted Least Squares Procedure



(1) Take s_u^2's as initial estimates of the γ_u^2's.

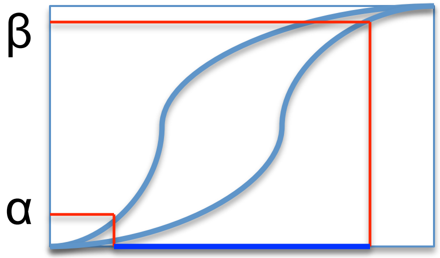
(2) Use weighted least squares on the $\bar{y}_u$'s to estimate the β's for the current values of the γ_u^2's.

(3) Reestimate the γ_u^2's using the current estimates of the β's.

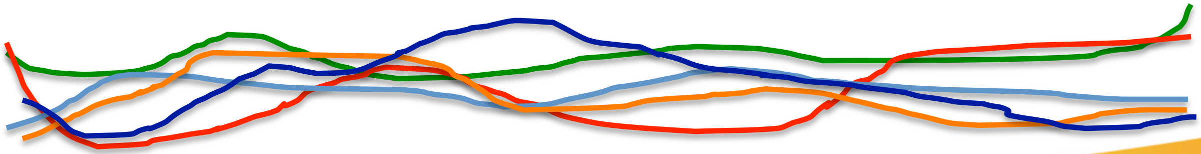
Iterate steps (2) and (3) until a convergence criterion is met.

Current Directions in Statistics and Uncertainty Quantification

- Computer experiments and the rise of UQ
- GASP Models and Discrepancy Analysis
- Uncertainty and Exascale Computing, propagating uncertainty through workflows
- Computing with Confidence using c-boxes and p-boxes (Ferson, 2013)



- Contour Boxplots, characterizing uncertainty in ensembles using data depth (Whittaker, 2013)



UNCLASSIFIED

References

- Bartlett, M. S., Kendall, D. G. (1946), “The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation,” *J. of the Royal Statistical Society,* Series B, 8, 128-150.
- Berkson, J. (1950), “Are There Two Regressions?” *J. of the American Statistical Association*, 45, 164-180.
- Box, G.E.P. (1979), “Robustness in the Strategy of Scientific Model Building, in *Robustness in Statistics*, ed. By R. L. Launer and G. N. Wilkinson.
- Box, G. E. P. (1963), “The Effects of Errors in the Factor Levels and Experiment Design,” *Technometrics*, 30, 1, 1-17, 38-40.
- Box, G. E. P, Fung, C. A. (1986), *Studies in Quality Improvement: MinimizingTransmitted Variation by Parameter Design, U. of Wisconsin Center for Quality and Productivity Improvement, Report No. 8*.
- Box, G. E. P., Cox, D. R. (1964), “An Analysis of Transformations,” *J. of the Royal Statistical Society*, Series B, 26, 2, 211-252.
- Deming, W. E. (1943), Statistical Adjustment of Data, reprinted in New York by Dover (1964), 37-48, 173-187.
- Dempster, A., Laird, N., Rubin, D. (1980) “Iteratively Reweighted List Squares for Linear Regression When Errors are Normal/Independent Distributed,” Multivariate Analysis, V, 35-37.
- Efron B, Hinkley, D. V. (1978), “Assuming the Accuracy of the Maximum Likelihood Estimator
- Ferson, S., Balch, M., Sent, K., Siegrist, J. (2013), “Computing with Confidence,” Proceedings of the 8th International Symposium on Imprecise Probability: Theories and Applications, Ed. By F. Cozman, T. Denoeux, S. Destercke, and T. Seidenfeld. SIPTA, Compiegne, France.
- Ku, H.H. (1969), “Notes on the Use of Propagation of Error Formulas,” Precision Measurement and Calibration, Selected NBS Papers on Statistical Concepts and Procedures, ed. By H.H. Ku, Washington, DC: U.S. Government Printing Office.

UNCLASSIFIED

References

- Leon, R. V., Shoemaker, A. C., Kacker, R. N. (1987), “Performance Measures Independent of Adjustment, Technometrics, 29, 3, 253-285.
- Madansky, A. (1959), “The Fitting of Straight Lines When Both Variables Are Subject to Error,” *J. of the American Statistical Association*, 54,173-206.
- Morrison, S. J. (1957), “The Study of Variability in Engineering Design, *Applied Statistics*, VI, 2, 133-138.
- Taguchi, G., Wu, Y. (1985), *Introduction to Off-Line Quality Control*, Nagaya, Japan: Central Japan Quality Control Association.
- Weigle, J. C. (2006), *The Effects of Temperature and Humidity on Wilethane 44*, LA-UR-06-3833, Technical Report, Los Alamos National Laboratory.
- Wendelberger, J.R. (2015), “Variation in Controlled Experimental Variables,” to appear in *Quality Technology and Quantitative Management*.
- Wendelberger, J. R. (2010), “Uncertainty in Designed Experiments,” *Quality Engineering*, 22, 88-102.
- Wendelberger, J. R. (1991), *Impact, Identification, and Estimation of Sources of Transmitted Variation*, Ph.D. Thesis, U. of Wisconsin-Madison.
- Wendelberger, J. R., and Box, G. E. P. (1992), “Identification and Estimation of Sources of Transmitted Variation,” *Proceedings of the 1992 American Statistical Association Meeting, Section on Physical and Engineering Sciences*.
- Wendelberger, J. R. (1993), “Efficiency Effects of Variation in Controlled Experimental Variables. *Proceedings of the 1993 American Statistical Association Meeting, Section on Physical and Engineering Sciences*.
- Whittaker, R., Mirzargar, M., and Kirby, R. (2013), “Contour Boxplots: A Method for Characterizing Uncertainty in Feature Sets from Simulation Ensembles,” *IEEE Trans. Vis. Comput. Graph,* 2713-2722.