Project Title: Wheat-field Variation and Nitrogen Fertilization Practices

Project Leader: Michael E. Tarter, *Professor of Biostatistics,*
University of California, Berkeley
(510) 642-4601
tarter@berkeley.edu

Cooperators: Steve Wright, *Farm Advisor –*
Tulare/Kings Counties, UCCE Tulare County
4437 S. Laspina, Ste. B. Tulare, CA 93274-9539, 559-684-3315,
sdwright@ucdavis.edu **and,**

Steve Orloff, *Farm Advisor/ County Director,*
UCCE Siskiyou County, 1655 S. Main St. Yreka, CA
*(530) 842-2711, sborloff@ucdavis.edu***:**

Abstract/Summary of Results and Conclusions:
Based on data supplied by the Cooperators a paper was presented this July at the San Diego International Joint Statistical Meetings. The following two findings were presented:

(1) Regarding wheat-field variation, yield protein variate, Y_{wp}, has an estimated density that is asymmetric and,
(2) For a sample of the Y_{wp} variate's measurements, a specialized logarithmic transformation can be determined that, after variate transformation, yields the type of bell-shaped distribution that has been shown, in a 1969 paper (in Biometrika) to enhance correlation coefficient reliability. Findings 1 and 2 above concerned the dependent variate, Protein Yield. Regarding the dependent variate, N rate in lbs/acre, based on preliminary studies conducted during the first project year we now plan to do 80 different rates from 0 to 400 clustered around 275 to 325. The basic idea here is to concentrate the spacing between adjacent selected N rates closest together at the value 300 lbs/acre. Both to the left and to the right of the value 300 the spacings between adjacent rates will increase. The procedure for spacing optimization is designed to increase the precision of measured information near the value (here 300) deemed optimal based on prior Cooperator experience. The new approach will help assure that the assumptions which underlie both linear regression and also Pearson correlation methodology are satisfied.

Introduction and Objectives:
Unlike many agronomy studies, our investigations concern the correlations between measurements, not contrasts and comparisons between subgroups of measurements. Technically speaking, interrelationships are studied by means of multivariate analyses while often contrasts and comparisons follow univariate Analyses of Variance and Covariance. The assumption that variates have a joint multivariate normal, bell-shaped distribution—if it is a valid assumption—expedites the study of interrelationships. (The equation for a special case of the multivariate normal model is provided by Expression (4) of the appended preprint.) In some circumstances a

newer model-free (nonparametric) technique, one that, in general, is likely to be substantially more robust than its parametric counterpart, can be applied. However, many studies have now established that nonparametric techniques can be subject limitations attributable to what are now called, edge effects and/or multimodality. Hence the problems addressed by the project and research objectives focus on the gathering, screening and cleaning of observations in order to expedite (as well as increase the scope and statistical power) of wheat-related studies of measurement interrelationships.

Materials and Methods: Regarding multivariate analysis specifics, the issue of primary concern continues to be the determination of a sampling method (plant part and timing) that is most predictive of late season nitrogen needs to achieve high yield and a protein concentration that meets quality standards (14 percent for Northern California and 13 percent for Central California). With a tool that accurately indicates the nitrogen status of the field, we can determine if a late season fertilizer application is needed to make protein.

IERC Grain Harvest data pertaining to crops planted 4/29/11 was forwarded to the PI by the Cooperators. (Measurements were recorded 9/19/11.) It was from this forwarded data that results of the experiments conducted to accomplish project objectives; as well as the figures shown below were based. Regarding statistical methodology, an updated an entry of the Encyclopedia of Environmetrics (scheduled to appear early 2013) is attached at the end of this report. The last section of this encyclopedia entry describes the theory that underlies the project research results illustrated by the three figures shown in the Results Section.

Key findings and conclusions/recommendations: The practical significance is primarily that by using new statistical techniques we hope to identify the most predictive sampling technique. By doing so wheat producers can accurately assess whether a late season application of nitrogen is needed in order to meet protein requirements in the marketplace. Natural and simulated data trials indicate that the new method is both general and applicable to a wide variety of subject matter. Hence we recommend that in the second project year attention be turned to the issue of independent variable level selection with special emphasis on using prior information obtainable from the cooperators. Hopefully we will have as much success with independent variable level selection improvement as we have had, this last project year, with dependent variate transformation. Hence, although the statistical methods used to obtain research findings are, admittedly, complex, the findings themselves are likely to serve the interests of farmers and wheat industry representatives.

Budget: Weekly phone conferences with Steve Wright, Farm Advisor –Tulare/Kings Counties, and Steve Orloff, Farm Advisor/ County Director, Siskiyou County, laid the groundwork for a meeting with Steve Orloff and his staff held at the Intermountain Research and Extension Center, Tulelake Northern California on June 29, 2012. The next formal meeting, with Steve Wright's staff members (Lalo Branuelos and Sonia Rios) took place in Coalinga, Central California, July 26, 2012.

On July 31, 2012, preliminary project-related findings were presented as a contributed paper at the International Joint Statistical Meeting held in San Diego. (Following this presentation, the PI was asked by the 2013 Montreal International Joint Statistical Meeting, Risk Section Chairperson to present an invited paper entitled: "Malnutrition-Environmental Degradation, Risk Tradeoffs with Special Emphasis on Wheat Protein Forecasting.")

According to the UC Berkeley, Biostatistics Administrator, "Per your request, attached is a fund summary report for your CA Wheat project." The attached report reflects that your current balance is $2,448.91. However, that will reduce by $2,294.51 once your travel reimbursement from your recent trip to Coalinga/San Diego hits the ledger. Your ending balance will then be: $154.40." Below please find the current report referred to above:

Business Unit: 1 – UC Berkeley
Fiscal Year: 2012 - 13
Month: From July to September

CUSTOMER REPORT 9 col (summary)
By Dept ID

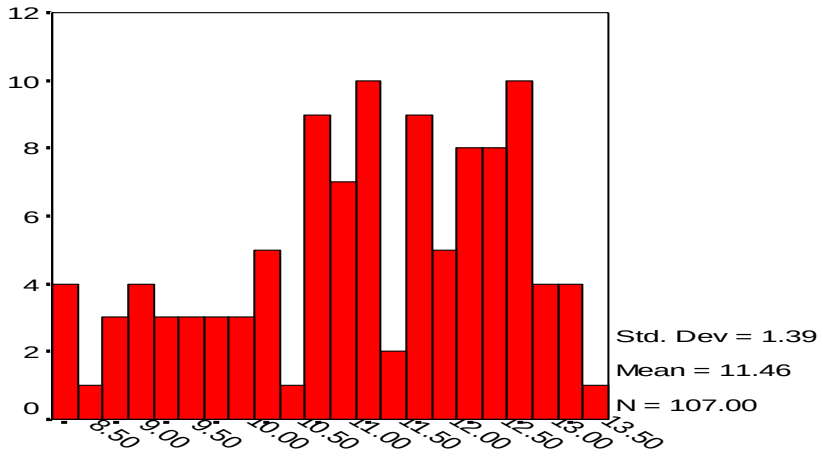
Page 1 of 1
Run Date: 09/27/12
Run Time: 11:22:52

Selection Criteria: Account Code | Fund Code 96299 | Dept ID 14043 | Program Code | Chartfield 1 | Chartfield 2 CPT89 Org L2 Node | Org L3 Node | Org L4 Node | Org L5 Node | Org L6 Node | Acct L2 Node: EXPENSES | Acct L3 Node: | Acct L4 Node:

Category Description	Prior Budget	Current Budget Activity	Total Budget	Prior Expenses	Current Expense Activity	Total Expenses	Encumbrance	Pre Enumb.	Balance
Dept ID 14043 - CPSPH BIOS Research /PI									
DIRECT COSTS									
Supplies & Expenses	0.00	0.00	0.00	2,343.27	453.75	2,797.02	0.00	0.00	2,797.02
Domestic Travel	0.00	0.00	0.00	0.00	754.07	754.07	0.00	0.00	754.07
Unallocated	-6,000.00	0.00	-6,000.00	0.00	0.00	0.00	0.00	0.00	-6,000.00
TOTAL DIRECT COSTS	-6,000.00	0.00	-6,000.00	2,343.27	1,207.82	3,551.09	0.00	0.00	-2,448.91
TOTAL FOR Dept ID 14043	-6,000.00	0.00	-6,000.00	2,343.27	1,207.82	3,551.09	0.00	0.00	-2,448.91

Results:

Many of the findings obtained with the support of project funding were motivated by protein yield findings described by Figure 1



1. PROTEIN

Figure 1. Histogram of Protein Percentage Variate

Notable is the seeming gap over the value 10.5% Protein. Hence there is a fundamental, often-encountered, question. Is there a lurking variable whose identity, if determined, would separate two groups of experimental units, here high-protein versus moderate-protein units? Alternately,

is the apparent dichotomy of protein level due to an artifact attributable to the statistical procedure used to display information?

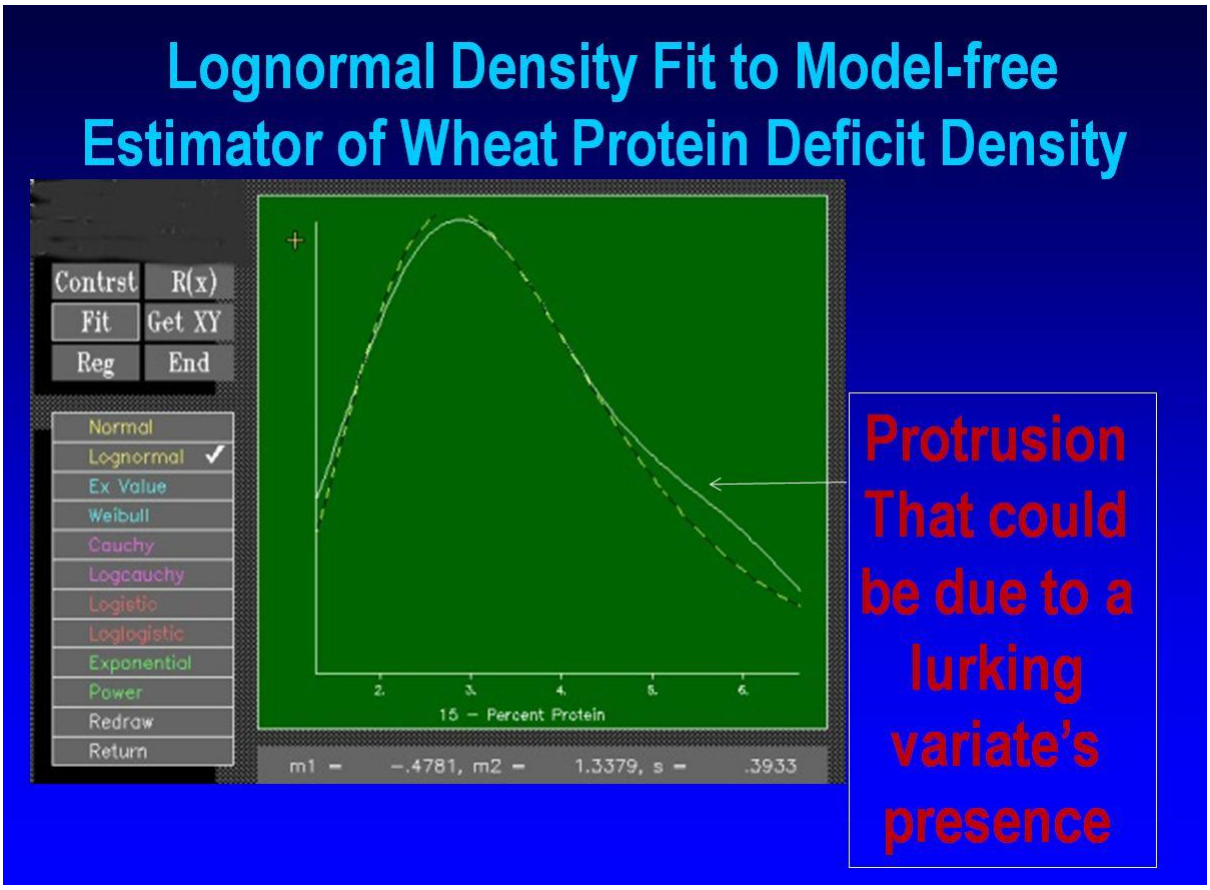


Figure 2. Histogram of 15% minus Protein Percentage Variate

Figure 2 Shows an important difference and an important feature similar to a feature observed the Figure 1 histogram. The difference is that since the lognormal, Chi-square, and most other asymmetric curves have an elongated right tail, rather than working with the protein percentage variate directly, a protein-deficit variate, specifically, 15% minus measured protein, is graphed along the x-axis. However, similar to the Figure 1 histogram, the protrusion or bulge shown to the left of protein-deficit level 5, suggests that there is a mixture of two distinctly different statistical densities each attributable to a distinct type of planted wheat.

There have been many books, papers and monographs that have been written to clarify the issue raised by Figures 1&2's bump, protrusion and/or gap between population subgroupings (http://en.wikipedia.org/wiki/Mixture_model). Hence, in the context of wheat protein studies, a useful first step was to construct an easily modifiable program that would help establish whether there were really two distinct subgroups or, alternatively, whether apparent bifurcation was attributable to a methodological artifact. The output of a new computer program, one designed with an eye to the bifurcation issue, is shown by Figure 3.

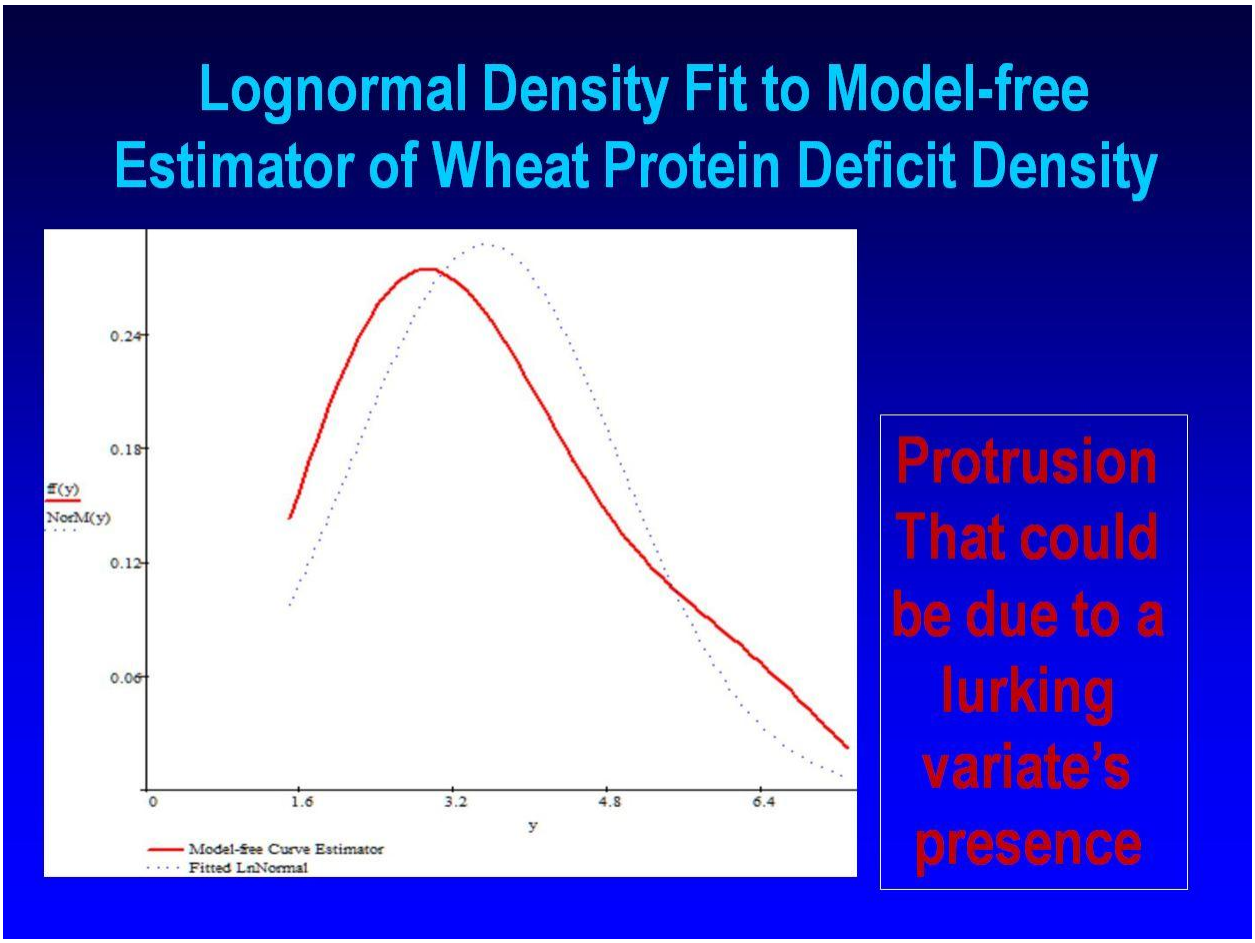
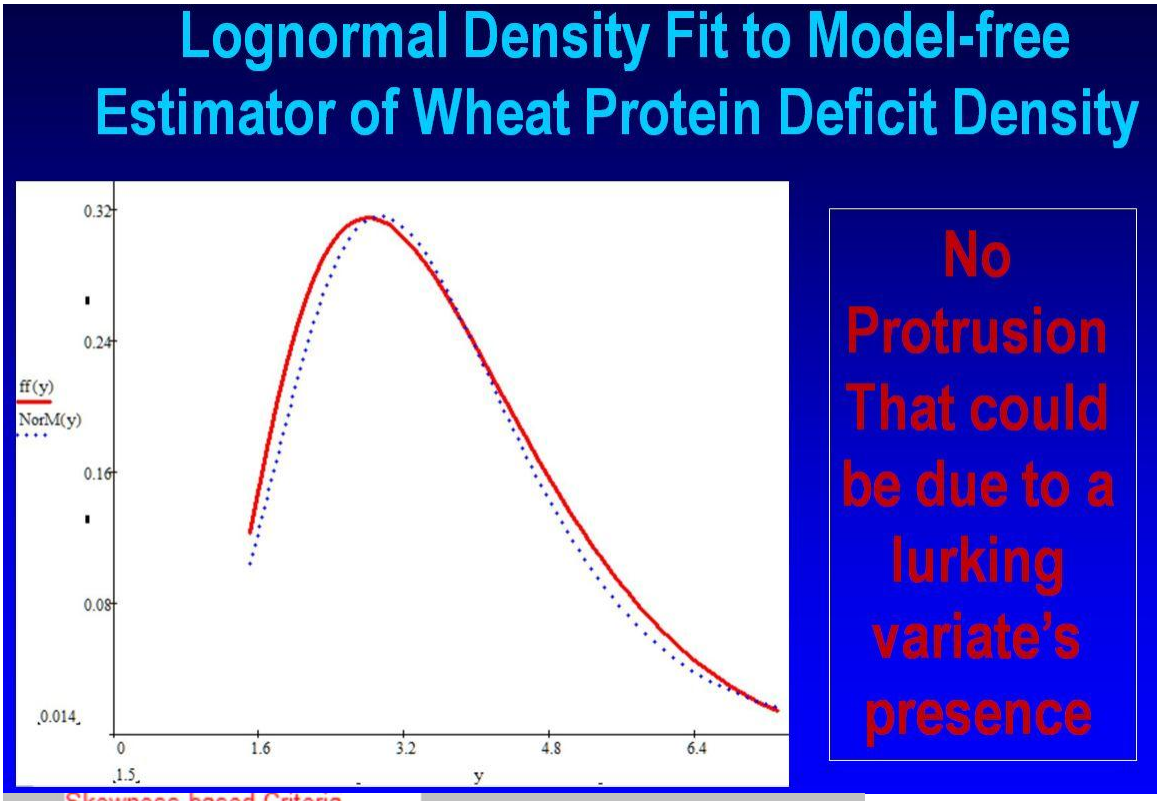


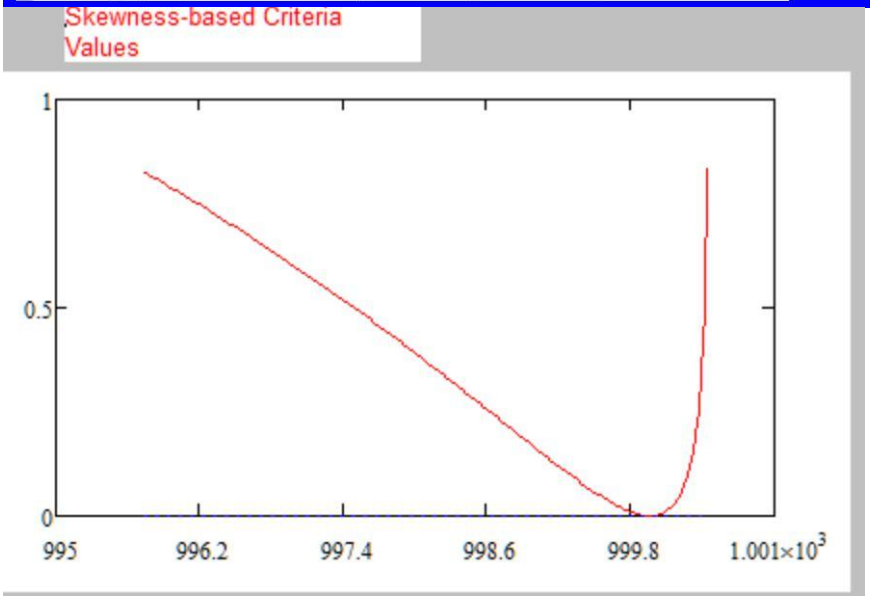
Figure 3. MATHCAD program designed to match methods used to obtain Figure 2.

One potential cause of apparent bifurcation is the need for a data transformation of the observed variate, X', to the new variate, X = ln(X' - τ). Section 2 of http://www.dtic.mil/cgi-bin/GetTRDoc?AD=AD0647325, as well as many other publications describe alternative approaches for estimating and applying estimates of τ. The problem is, most of the literature on this subject assumes that X will have a well-researched density such as the normal or logistic frequency function. However, the procedures whose implementation helped graph Figures 2&3 are all designed to be model-free, i.e. nonparametric. Hence the novel feature of the estimator

$TAUF = -0.966$ (an estimator obtained by applying project-related findings described in the figures and table shown below) is that its determination did not rely on the assumption that post-transformation variate X has a density that can be modeled by an elementary function such as the normal or logistic density model.

The figures shown below illustrate both how Figure 3 was modified by means of a pilot estimate of τ and, how extensive trials with simulated data where τ was set equal to the value 1000 yielded an estimate, here the point on the axis where skewness assumed its lowest value.





Figures 4&5. Natural data and simulated data estimated τ-based findings.

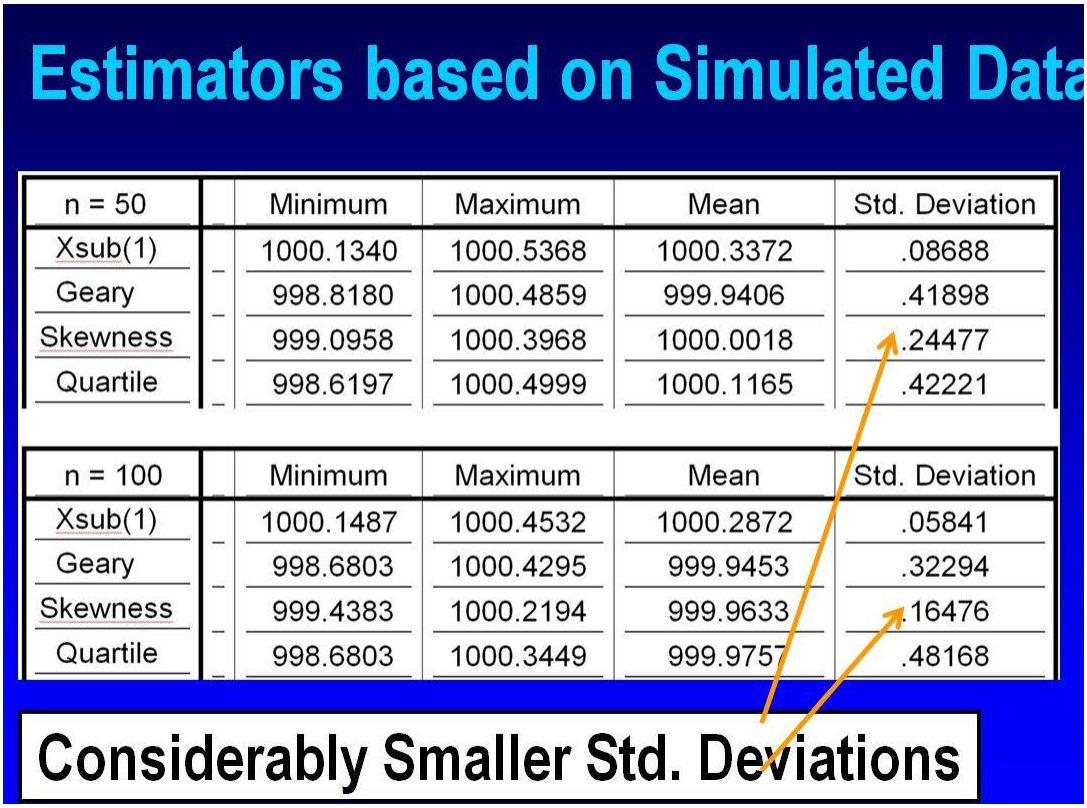


Table 1. One of the many tables constructed in order to obtain a useful model-free estimator of the threshold parameter.

Discussion, Conclusions and Recommendations: Based on data supplied by the Cooperators, a paper at this year's International Joint Statistical Meetings, San Diego, presented the following two findings: (1) Regarding wheat-field variation, yield protein variate, Y_{wp} , has an estimated density that is asymmetric and, (2) for a sample of the Y_{wp} variate's measurements, one specialized logarithmic transformation can be determined that, after variate transformation, yields the type of bell-shaped distribution that has been shown to enhance correlation coefficient reliability.

Implications of the results of the research on project objectives.
While in the first project year we have implemented a statistical procedure that can screen and clean dependent variate measurements in our second project year we hope to be able to allocate N rate in lbs/acre so that multivariate approaches can be applied effectively. (Initially we plan to do 80 different rates from 0 to 400 clustered around 275 to 325.) The basic idea here is to concentrate the spacing between adjacent selected N rates closest together at the value 300 lbs/acre. Both to the left and to the right of the value 300 the spacings between adjacent rates will increase. The procedure for spacing optimization is designed to increase the precision of measured information near the value (here 300) deemed optimal based on prior Cooperator experience. As is certainly true for the procedure described by the figures and table above, the new approach will help assure that the assumptions which underlie both linear regression and Pearson correlation methodology are satisfied.