# NEW TOOLS FOR NEW TIMES

Terry C. Nelsen

Debra E. Palmquist

**Kansas State University**

NEW TOOLS FOR NEW TIMES

Terry C. Nelsen and Debra E. Palmquist,
USDA-ARS-MWA, 1815 North University Street, Peoria, IL 61604

Abstract

The purpose of this presentation is to challenge statisticians to develop new tools needed by modern scientists. We are in the midst of a Scientific Revolution being driven by computers and the internet. Scientists are gathering huge amounts of data on the usual measurements while continually developing new instruments for new measurements. Data sets full of measurements which may pertain to the scientist's research are easily available on the internet. Scientists are being overwhelmed with data. Agricultural producers and consumers are asking for more information. Scientists need new tools to evaluate variation. They need help with sampling – numbers of observations required and proper sampling schemes. Examples and suggestions will be offered. Statistical Process Control as applied to farming systems will be discussed.

Key Words: Statistics, Internet, Decision Support, Agriculture, Measurement, Weather, Precision Farming, Database

1. Introduction

For a century or more, statisticians have provided tools for agricultural researchers to design their studies and then analyze and interpret their data. Student's t-test was developed to test agricultural commodities being delivered to breweries (Student ,1923). The first published Analysis of Variance was used to compare the response of different varieties of potatoes to different levels of manure in the presence of variation in the soil (Fisher and MacKenzie, 1923). In the 1950's C. R. Henderson conceived BLUP (Best Linear Unbiased Prediction), for the resolution of equations in mixed models used for the genetic evaluation of animals (Henderson, 1953). Geneticists have used these mixed-model techniques to make wonderful improvements in plants and animals. Modern bio-informaticists are helping molecular biologists interpret the genomic and proteomic data being produced at increasing rates at labs all over the world.

The world of research is in the midst of a revolution. Modern electronic instruments are gathering data faster and more accurately. These instruments are collecting data on different measurements. All of this new data on traditional and new measurements means that scientists require new and efficient tools from the statisticians supporting them. New experimental designs and analytical tools will advance the revolution.

2. Decision Support Tools

A farmer starts out with soils of specific types that have been conditioned by past practices and crops. Fields can be tilled and prepared at varying times of the year and with different equipment types and settings. The farmer decides which crops to plant and then which fertilizers and pesticides will be applied, how they will be applied, at what rate, and when. The environment provides variation in temperature, wind, humidity, cloud cover, moisture and storms. Pests, diseases and weeds challenge the crop. Each of these input variables can be evaluated as to their effects on responses such as Yield, Economic Return on Investment, Environmental Factors – Air, Water, and Soil Quality, and even Social Factors, such as personal satisfaction and neighbors' concerns.

All of these factors are difficult to consider simultaneously. There are scale effects – a study done on a plot or subplot can be quite different from a similar study done on a field or watershed scale. Multiple interactions complicate the analysis. The task may look too complicated, but modern computers and forward thinking statisticians are able to use modeling techniques to increase the farmers' chances of knowing what and where they need to apply or change their management practices.

Farmers have asked us to provide them with information which they can use to make their individual decisions. We have provided many tools already and the demand grows. Jame and Cutforth (1996) described many of the decision support models currently available. One of the earliest and most successful has been a model named GOSSYM which is capable of making site specific recommendations to cotton farmers for seeding, nitrogen applications, plant growth regulator applications, and insect control (Hodges et al., 1998). The Nebraska Planting Guide (NRI-Nebraska, 2003) provides information to guide Nebraska farmers in selecting planting dates for different corn and soybean varieties. The Cooperative Extension System and Land Grant Universities have formed an on-line ADDS Center (ADDS = Agricultural Databases for Decision Support, at http://www.adds.org/about.asp) to compile extension materials, university research reports and private industry documents into easily accessible, readily available reference materials because "Agricultural producers are faced with ever increasing demands from an economic, environmental and social standpoint. Complicating matters is the increasing complexity of technologies and management systems that require far-ranging expertise and rapid availability of information. Decisions that were once made with confidence by the manager using his or her own expertise now often require information and expertise from other resources. Decisions made in the privacy of the farm operation may now have significant implications beyond the farm gate."

3. The Measurement Revolution

Modern electronic instruments are rapidly changing what scientists measure as well as how accurately, how often, and at what cost. We have measured protein in plants and animal tissue for many years. The long-time standard was the Kjeldahl Method, which took about two hours per group of samples. In the 1990's a combustion method (AOAC, 1990) was adopted. The combustion method takes about five minutes and is safer and more accurate. Currently a Near Infrared Reflectance (NIR) method (Kays et al., 2000) is being used which is portable, nearly instantaneous and does not destroy the sample. Companies such as NDC Infrared Engineering will supply an instrument which can measure moisture, fat, oil, sugar, and protein in foods continuously and non-destructively on a moving assembly line (NDC, 2003).

Just for a moment, compare the word "protein" with the word "plant". We all know what a plant is and we all can appreciate how many very different plants there are in the world. There are just as many differences in proteins. Proteins can make up as much as 50% of the dry weight of a cell. The many different proteins have different sizes, different shapes and different functions. For years, scientists considered protein as one entity in a cell and calculated "crude protein" in meat or in plants used for human or animal nutrition. Modern instrumentation and modern ideas are allowing scientists to differentiate individual proteins that can function individually or in combinations as enzymes, building blocks, storage sites, and especially as the messengers of the genetic code. Scientists are starting to acquire tools to help find, describe and categorize all of these proteins (Laskowski, 1996).

As another example, I foresee the term "Soil Organic Matter" quickly going the way of "Crude Protein" as better measurements become available.

Experimental Design for most agricultural research has focused on hypothesis testing and Type I errors. These designs have served us well. Means comparisons tests allowed agricultural producers to select breeds, varieties, fertilizers, pesticides, etc., to meet individual needs. For example, the Ohio State University Extension Service provides growers with information on-line for many currently available varieties of corn hybrids. They list means and an l.s.d. for yield potential, standability, maturity, and other agronomic characteristics that affect profitable crop production (Minyo et. al., 2002). I suggest that innovative statisticians can provide growers with more efficient methods of variety evaluations than simple means comparisons.

These experimental designs usually assume normality and homogeneous variances. These assumptions are robust and have allowed advances in agricultural production. Remember, however, that these designs were developed before computers when time and calculator limitations restricted the number and size of possible calculations. Statisticians need to continue to provide researchers with new tools that take full advantage of the power of computers. Mixed-model analysis has been a step in the right direction.

We can help researchers to change some of the focus of their research from comparisons of means to more evaluation and comparison of variances. Two breeds or varieties can have similar means in yield or quality measures, but environmental or management stresses can affect them quite differently. Variance in crop yields or quality will cause variance in the grower's risk. Grain varieties which are more variable increase the cost of processing and can have undesirable variation in end-use qualities and customer acceptance. Variation within a commodity can cost processors in time, money and final quality and they are interested in research which can stabilize or minimize or, at least, predict variation.

## 4. Statistical Process Control for Agriculture

One area of new tools to consider for agricultural research is as some modification of Statistical Process Control (SPC). Engineers use SPC to characterize manufacturing processes and properties. Characterization is measured as central tendency and variability. Quality Control has been around for centuries in one form or another but *statistical* quality control is comparatively new. It was not until the 1920s that statistical theory began to be applied effectively to quality control as a result of the development of sampling theory. A good summary of the historical background of SPC is found in Chapter 1 of Quality Control and Industrial Statistics, by Acheson J. Duncan.

Farming and ranching are production systems that have much in common with manufacturing systems but are complicated by their inherent variability in biological, environmental, and management systems. Plants and animals with their own genetic variability grow and produce in variable soils in variable climates under variable management conditions as they are exposed to different pests, diseases and environmental challenges. Can we provide tools to agricultural researchers that examine farming systems to determine the critical points causing variation in livestock or plant production?

5. Precision Farming and On-Farm Research

Our attempts to conduct on-farm large scale research in the past have been limited by the difficulties of superimposing our traditional experimental designs on the day-to-day operations of a farm or ranch. We have had difficulties in controlling variation due to soil differences. Some of Fisher's (Fisher, 1920) early work was to control the effects of soil variation. Farmers and researchers now have the tools to collect data on smaller and smaller subplots within farm fields. Modern GIS systems can provide locational accuracy at the centimeter level and are being combined with instruments installed on a harvester for continuous on-the-go analysis of grain protein as well as other quality characteristics (Jensen, 2001). The United Soybean Board funded a project (USB Project #2209) in which 16 researchers in 15 states were asked to "explore questions concerning site-specific management and the use of data in making farm management decisions." One of the problems with Monte Carlo studies has been that we were not sure of the underlying variation. We now have the information to determine the actual variation in many different measurements within a plot or field.

6. How Many Measurements?

Scientists often ask us "how many measurements do I need?" We have approached this question from the standpoint that measurements are expensive in terms of time, space or cost and the scientist wants to know the minimum number of data points required. Modern scientists are using new electronic instruments and internet-based data sets and they are being overwhelmed with data. Scientists have access to precipitation data reported 4 times per hour for over 3400 U.S. stations collected since 1971 (NCDC, 2003). In Ohio, (OARDC1, 2003) data is available on weather data at 15-minute increments at 15 locations around the state. Does a grower need temperature every 15 minutes to make an accurate prediction? How many measurements does it take? Also, is data from a weather station 10 miles away adequate?

I am aware of a study (Prueger et al., 2003) which is designed to compare on-the-ground measurements of heat, water vapor, and carbon dioxide with remotely sensed data from an aircraft. The instruments are capable of recording 20 times per second on six or more variables and in 20 days have produced 415 million lines of data (Prueger, personal communication).

In my experience as a consulting statistician for researchers in the USDA Agricultural Research Service, I found that sampling is a serious problem for researchers and complicates the interpretation of their data. We tell scientists to take random samples, but I have found that true randomness can be a difficult achievement. Confusion commonly arises when the differences between experimental units and sampling units are not clear. The scientists feel more comfortable with a stratified sample of some type because they worry that a true random sample has a probability of concentrating at one extreme or the other of a given measurement. Can we take advantage of some of these large data sets to do "what if?" studies? Start with the entire data set and show what would have happened if only one-half or one-third of the data had been collected. What would have been the consequences of random versus stratified sampling? We can talk about theory and Monte Carlo studies, but these large data sets can be used to quickly check the Monte Carlo theories.

7. Weather

Another example of an area where statisticians can help is with weather data. Currently, nursery and horticultural producers are using weather data to make decisions on management practices based on their concern about the possibilities of economic loss. They can contract with private weather services or use public services such as the Growing Degree Days and Phenology for Ohio Project (OARDC2, 2003). This project supports an interactive web-site that enables a producer anywhere in Ohio to input a 5 –digit zip code and obtain information as to which insect/pest the producer needs to spray for and when. The goal of this project is to reduce the total number of sprayings. It is currently common practice for growers to spray their nurseries every five days. This web-site provides a list of the phenological events as and when they occur, so that better pest management decisions can be made.

There are data sets available on the internet with years of data on weather and soil conditions. Some of these data sets are huge, however. The Natural Resources Conservation Service's Centralized Database System (CDBS, 2003) has data from nearly 17,000 climate stations, of which 6,700 contain greater than 20 years of data. The World Data Center (WDC, 2003) for Meteorology in Asheville, NC, is just one component of a global network of centers that facilitate international exchange of scientific data. Their mission is to acquire, catalogue, and archive data and make them available to requesters in the international scientific community.

Can we can use weather data from growers' on-site weather stations or from nearby National Weather Service sites along with on-line databases to instantly and continually predict weed infestations, insect hatches, the onset of different diseases, or crop maturity? We can help the researcher to learn the effects of soil temperature (absolute, minimum or maximum, or duration of minimum or maximum), moisture, soil organic matter, sunlight, slope (direction or steepness), or the presence or absence of natural or introduced chemicals. I would expect these variables to be interrelated and the scientist will need some sophisticated methods of analysis to find answers. Until the development of modern computers, developing, testing and using sophisticated methods could be tedious. We are no longer restricted by calculation intensity and should not be restricted by its residue mindset.

8. Spatial Analysis

Many designs are available for spatial analysis. Most of these designs were developed by the mining or petroleum industries where conditions are relatively static. Agricultural scientists need designs to study spatial variation which can change as plants mature or rains occur. We need designs for spatial/temporal variation.

At what scale should data be sampled? In the previously mentioned soybean study (USB, 2002), some interesting approaches to resolving scale differences have been published (Brouder, 2002). The Geographic Information Retrieval and Analysis System (Steeves and Nebert, 1994) divides the U.S. into Hydrologic Unit Codes (HUCs). The hydrologic units are encoded with an eight-digit number that indicates the hydrologic region (first two digits), hydrologic subregion (second two digits), accounting unit (third two digits), and cataloging unit (fourth two digits). Hydrologists discuss doing research at an eight digit scale, a six digit scale, etc.

9. The Internet Revolution

Think of the internet as the modern version of the printing press. The printing press changed society. The ability to print books changed the centers of knowledge away from monasteries where monks labored over reproducing books. Philosophers, artists, scientists, and all types of thinkers and doers were no longer restricted in what they could publish but could print and disseminate their thoughts and findings to a wide audience. Centers of knowledge moved away from the monasteries and into the universities. At the other end, readers of books had much more access to books and pamphlets. There was more incentive for more people to learn to read. Languages had to become more standardized as books became more widespread. The internet may be moving Centers of Knowledge away from the Universities and into cyberspace. More citizens have access to more information and are quickly acquiring computer skills to use that access. Whether more information translates into more knowledge for the average citizen is a question for the philosophers. We can and should help the scientists translate that information into knowledge. Scientists will be using the internet to gain access to data pertinent to their experiments. For example, instead of collecting their own weather and soil data for each experiment, they perhaps can find a database on the internet which might give them the information they need. They need the tools to manipulate that raw data into a form useful to a particular experiment.

The National Water Information System Web Site (NWISWEB, 2003) contains access to water-resources data collected at approximately 1.5 million sites in all 50 States. The files are listed for water flow and levels along with chemical and physical data for wells, streams, lakes and rivers. The National Resources Inventory (NRI, 2003) has data on land cover and use, soil erosion, prime farmland soils, wetlands, habitat diversity, selected conservation practices, and related resource attributes collected every 5 years from the same 800,000 sample sites in the U.S. The NRI is a statistically based survey that has been designed and implemented using scientific principles to assess conditions and trends of soil, water, and related resources.

I foresee databases on soils, weather, pests, diseases, moisture, etc. and the effects of different management practices being available for any given location in North America. Forward thinking scientists will want to use all available resources for their research but they will need tools to determine the proper experimental designs which take advantage of all of these resources and then more tools to determine which data in what form is most appropriate.

10. Summary

The computer age is upon us and our personal and professional lives are changing rapidly. I challenge research statisticians to work with researchers and provide them with new tools to take advantage of all of the new resources available. One note of caution, however, is to listen carefully to the scientists to find out what tools they want and need. An axiom of business is to first identify your customers and then find out what they want. We shouldn't develop tools that impress other statisticians but can't or won't be used by agricultural researchers because they either don't understand the underlying assumptions or can't interpret the results. We can anticipate that tools which are developed to take advantage of the internet will themselves be available on the internet and should, therefore, be rapidly and widely dispersed among the scientists who use them.

Agricultural scientists in the midst of this scientific revolution are in great need of support from innovative statisticians. I have a perception of a growing gap between basic and applied research in the statistics discipline. In my opinion, this gap is being driven in part by the movement of university statistics departments towards more rigorous, basic research. I will never argue that any scientific field can ignore basics and survive, but I will argue that there are too many statistical scientists doing basic research and too few developing new tools for applications. The statistics departments in some of the Land Grant Universities are losing sight of their original reason for being.

This separation of academia and application is not unique among statisticians. A group of the leading engineering schools in the US and Europe have formed a collaborative (CDIO Initiative, 2003) to address their concerns about the widening gap between engineering education and real-world demands on engineers. The CDIO stands for Conceiving — Designing — Implementing — Operating process and students are being asked to learn to apply their education on practical problems simultaneously with their formal education. Perhaps statistical educators should consider a similar initiative.

## References

ADDS Center, Agricultural Databases for Decision Support. 2003. Available at: http://www.adds.org/about.asp, accessed October 22, 2003.

AOAC, Association of Official Analytical Chemists. 1990. Official methods for plant analysis. 15th ed. Kenneth Helrich (ed.) AOAC, Arlington, VA.

Brouder, Silvie M. 2002. Evaluation of Site-Specific Management for Indiana Soybean Production Systems. available at: http://www.farmresearch.com/usb/Midwest/IN16/in16Reports/2002Final-IN.pdf, accessed October 28, 2003.

CDBS, Natural Resource Conservation Service, US Department of Agriculture, Centralized Database System. 2003. Available at: http://www.wcc.nrcs.usda.gov/climate/ accessed October 27, 2003.

CDIO Initiative. 2003. Conceiving — Designing — Implementing — Operating real-world systems and products. Available at http://www.cdio.org/index.html, accessed October 28, 2003.

Duncan, Acheson J. 1986. Quality Control and Industrial Statistics, Irwin/McGraw-Hill, Chicago.

Fisher, R.A. 1920. A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error and by the Mean Square Error. Monthly Notices of the Royal Astronomical Society 80: 758-770.

Fisher, R.A., and W.A. MacKenzie. 1923. "Studies in Crop Variation: The Manurial Response of Different Potato Varieties," Journal of Agricultural Sciences 13: 311-320.

Henderson C. R. 1953. Estimation of variance and covariance components. Biometrics 9: 226-256.

Hodges, H.F., F.D. Whisler, S.M. Bridges, K.R. Reddy and J.M. McKinion. 1998. Simulation in crop management GOSSYM/COMAX. In:Agricultural Systems Modeling and

Simulation. R.M. Peart and R.B. Curry (eds.), Marcel Dekker, Inc., New York. Chapter 8, pp. 47.

Jame, Y.W and H.W. Cutforth. 1996. Crop Growth Models for Decision Support Systems. Can. J. Plant Sci. 76: 9-19.

Jensen, T., L. Zeller and R. Kelly. 2001. Mechanising on-the-run grain sampling. Proceedings of the 10th Australian Agronomy Conference, Hobart, 2001, available at: http://www.regional.org.au/au/asa/2001/1/a/jensen.html, accessed October 23, 2003.

Kays, Sandra E., F. E. Barton, II and W. R. Windham. 2000. Predicting protein content by near infrared reflectance spectroscopy in diverse cereal food products. Near Infrared Spectroscopy 35–43.

Laskowski, R.A, J.A.C. Rullmann, M.W. MacArthur, R. Kaptein and J.M. Thornton. 1996. AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. Journal of Biomolecular NMR 8: 477-486.

Minyo, R.J., Jr., A.B. Geyer, P.R. Thomison, B.L. Bishop and D.G. Lohnes. 2002. Ohio Corn Performance Test. Ohio State University Extension Ohio Agricultural Research and Development Center, The Ohio State University. Horticulture and Crop Science Series 215 November 2002, available at: http://www.oardc.ohio-state.edu/corn2002/#procedures, accessed October 27, 2003.

NCDC. 2003. National Climatic Data Center, National Oceanic and Atmospheric Administration, available at: http://cdo.ncdc.noaa.gov/CDO/dataproduct, accessed October 28, 2003.

NDC. 2003. NDC Infrared Engineering, A Spectris Company, available at: http://www.ndcinfrared.com/TechnologyView.aspx?technologyid=5&grouptypeid=5, accessed October 29, 2003.

NWISWEB. 2003. US Geological Survey, National Water Information System Web Site, available at: http://waterdata.usgs.gov/nwis/, accessed October 28, 2003.

NRI . 2003. National Resources Inventory (NRI) , Natural Resource Conservation Service, US Department of Agriculture, available at: http://www.nrcs.usda.gov/technical/NRI/1997/docs/background.html, accessed October 28, 2003.

NRI –Nebraska. 2003. Nebraska Research Initiative. Planting Date Guide, available at: http://www.plantguide.com/plantdatefinal/index.html, accessed October 28, 2003.

OARDC1. 2003. Ohio Agricultural Research and Development Center, Wooster, Ohio. OARDC weather network, available at: http://www.oardc.ohio-state.edu/centernet/weather.html, accessed October 28, 2003.

OARDC2. 2003. Ohio Agricultural Research and Development Center, Wooster, Ohio. Growing Degree Days and Phenology for Ohio Project, available at: http://www.oardc.ohio-state.edu/gdd/Default.htm, accessed October 28, 2003.

Prueger, J.H., W.P. Kustas, L.E. Hipps, J.L. Hatfield, A. Cahill, C. Williams, J.I. MacPherson, M. Wolde, J. Albertson, W.E. Eichinger, D.I. Cooper, N. Brunsell and R. Gillies. 2003. Spatial Variability of Turbulent Fluxes Across a Corn/Soybean Production Region in Central Iowa. Proceedings, 83rd Annual Meeting, American Meteorological Society. Details of the experiment are at: http://hydrolab.arsusda.gov/smex02/, accessed October 29, 2003.

Steeves, P. and D. Nebert. 1994. Hydrologic units maps of the Conterminous United States, available at: http://water.usgs.gov/GIS/metadata/usgswrd/huc250k.html#Distribution_Information, accessed October 28, 2003.

Student. 1923. "On Testing Varieties of Cereal". Biometrika 15: 271-293.

USB. 2003. Project #2209: Site-Specific Soybean Systems: A Combined Multi-Regional Plan, available at: http://www.farmresearch.com/usb/default.asp, accessed October 27, 2003.

WDC. 2003. World Data Center for Meteorology, Asheville,National Climatic Data Center, available at: http://www.ncdc.noaa.gov/oa/wmo/wdcamet.html, accessed October 27, 2003.