Statistics Education and the Role of Technology

Gail Burrill
Mathematical Sciences Education Board, National Research Council, Washington, D.C.

The world is changing dramatically. CD's, the web, caller ID, grocery store debit machines, MRI's: life today is very different from life thirty years ago. Computer assisted design, spreadsheet cost analysis, simulations, instant communication across cities and nations: the business world is very different from the business world thirty years ago. Information technology is transforming the way we do business and the way we live, and all indications are that even more change is imminent. In education just as in society and in the workforce, technology has the potential to make a profound impact. It can change the nature of the content we teach and the way we teach that content. In statistics, that potential is well on its way to becoming a reality.

Although a Danish astronomer, Tyco Brahe, (1546-1601) was one of the first scientists to confront the problem of measurement error (Freedman et al, 1996), even in the early 1700's the nature of statistics was primarily qualitative probably due to the fact that there was very little quantitative data available (Ottaviani, 1989). Quantitative information became a reality in Hungary in 1784 with the first census; the first U.S. census was in 1790. This data was very difficult to manage, and by the late 1880's, a crisis had developed in the attempt to organize and process the large amounts of data collected through the U.S. census. While mathematicians and statisticians had continued over the years to develop statistics as a field, Herman Hollerith in response to this crisis produced a machine that would handle tabulations mechanically (Wallechinsky & Wallace, 1975). In 1896, he organized the Tabulating Machine Company which eventually led to International Business Machines, now known as IBM, and indirectly, had an enormous influence on the nature of statistics as it currently exists. Today, technology allows the collection of and access to enormous amounts of data and can quickly organize and analyze that data. This information can be used for a vast array of endeavors: analyzing data to understand market trends, building models to understand and predict behavior, coding the interior of the human body to allow analysis without invasion, simulating the behavior of a rover on the moon, or making decisions about the effectiveness of a new drug. As information usage grows, so does the increasing importance of statistics as a fundamental tool in helping people manage and make sense of this proliferation of information.

This paper will focus on statistics education from two perspectives: first, a discussion and some recommendations about the nature of the statistics curriculum in secondary schools (which could be generalized to an introductory statistics course) and second a discussion of the role of technology and its impact on that curriculum.

The Statistics Curriculum

According to David Moore, teaching statistics should focus on what the practice of statistics is about (Moore, 1997):

- producing data that provide clear answers to specific questions,
- reading data critically and with comprehension, and
- using sound methods to draw trustworthy conclusions based on data.

In addition to the goals Moore lays out, the actual process of teaching statistics should be embedded in guiding principles relating the study of statistics and statistical reasoning to the practice of statistics. The Guidelines for Teaching Statistics (Burrill, 1991) recommended among other things that technology should be used to facilitate analysis and interpretation. Students should

- understand the need for and be able to formulate well-defined questions,
- begin with a graph,

- get involved with the data,
- focus on the big ideas not the rules,
- understand the value of different approaches and techniques,
- explore and experiment before using formal algorithms,
- experience the importance of randomization,
- identify assumptions,
- compare observed outcomes to chance outcomes,
- think critically and learn to ask the right questions.

Ideally, students should leave elementary school with some basic statistical knowledge: understand how to make and interpret bar graphs, number line plots, stem-and-leaf plots, and scatter plots; be familiar with median and mode as measures of center and range and quartiles as measures of spread; be able to relate measures of center and spread to the shape of the distribution. They should have some experience designing simple surveys and experiments. They should be able to make conjectures based on their data and provide supporting evidence for claims they make. While student learning should be grounded in hands-on experiences and pencil and paper techniques, there is increasing evidence that some computer graphics and drawing programs can effectively facilitate student understanding even at this level.

In lower secondary school, students should produce data by designing simple experiments and sample surveys, begin to appreciate the role of chance and to understand sources of bias and its impact on results. Simulation with concrete objects and simple technology applications can help students lay the foundation for more sophisticated thinking. Students should work with observational data collected from their environment, recognizing the limitations and constraints inherent in data gathered in this way. Headlines about a disease caused by mosquitoes and how scientists design experiments to track its spread or about the results of a study of the relation between day care and student achievement in school can help students understand the difference between observational studies and experiments. A poll about potential support for a new sports arena can lead to class discussions on question wording or the impact on the results of potential versus actual voters.

Students should explore the relationship between the shape, center, and spread of a distribution, be introduced to the mean and learn to summarize center and spread using the mean, quartiles, interquartile range, and outliers as well as the median and range they studied earlier. The graphical representations they use should expand to include box plots and histograms, and they should be able to compare data sets and describe their conclusions in words that make sense. They should understand linearity as a summary for bivariate data, be able to interpret the relationship between variables represented by a linear equation, and to describe the association as positive, negative or zero. Students should learn to read and interpret data from two-way tables. The notion of random samples can be introduced, and students can use simulation to begin to explore sampling distributions and their statistics. In this context they should recognize that random behavior has certain predictable characteristics and that the long run relative frequency of an event stabilizes. Technology can be an important tool in making these points meaningful for students.

In upper secondary school or an introductory statistics course, students should formalize their knowledge of collecting data through observations, sample surveys and experiments. They should understand the importance of well designed sample surveys and experiments and the sources of potential bias in their design and administration. This means they recognize terms such as control groups and blinding, and learn to critique media reports about experiments and surveys with a statistical framework. They should understand the role of randomness in reducing bias and recognize the effect of sample size. The data sets with which they work should be about subjects that are not restricted to their immediate environment. Their data exploration tools should expand to include standard deviation as a measure of spread, techniques to compare distributions, and a variety of models to represent the relationship between two data sets, including an understanding of residuals and of what correlation represents and does not represent. They should recognize when and how to use transformations to make sense of data presented in different scales or as a way to manage data of very different magnitudes and understand the impact of the transformation on the distribution and summary statistics. As citizens, students need to understand that using data and statistical reasoning is a legitimate and functional way to make a decision – for example, that sampling techniques can obtain a more accurate count of the population than attempting to count everyone. Students should be able to use statistical inference to support conclusions drawn from experiments or surveys; critique reports about sample surveys and experiments in the media; and use probability as a tool in making inferences. Students should be familiar with standard distributions such as the normal, the binomial, and the chi square and know how to compare an observed outcome to them.

Software packages such as Fathom from Key Curriculum Press, a dynamic computer learning environment for data analysis and statistics, can radically alter the approach to teaching these fundamental statistical concepts. Such software has capabilities that include dynamic manipulation of symbols, words, categories, and plots; dragging points, axes, attributes, bars, with continual updates; formulas attached to sliders and graphs for plotting functions, attribute definitions and use as filters; sampling (including random and selected cases) and collecting measures; and importing data directly from the internet. The power of such packages is not in doing statistics but in presenting concepts and engaging students in understanding the essential structures of the subject.

These guidelines are intended for all students to enable them to be statistically literate citizens as well as provide a statistical foundation for whatever future options they choose – continued study or the workplace. In the United States there is another option for those who are interested in university work, an Advanced Placement Statistics course designed to be taken while in secondary school but equivalent to a university course. If students score at a certain level on a national test, they can earn university recognition ranging from credit to advanced placement. A major difference between the curriculum for all and for those in advanced placement is the inclusion of formal inference -confidence intervals and tests of significance (College Board, 1999). As a side note, the first test in 1997 was given to about 7,500 high school students; 15,000 students took the test in 1998, and in 1999 the number topped 25, 000. Currently more and more schools across the nation are offering statistics courses to qualify their students for advanced placement. The students in these courses vary from those who like mathematics and take statistics because it is related to mathematics to those who are going into a liberal arts program and do not want to take calculus to those who need a credit to finish their high school coursework.

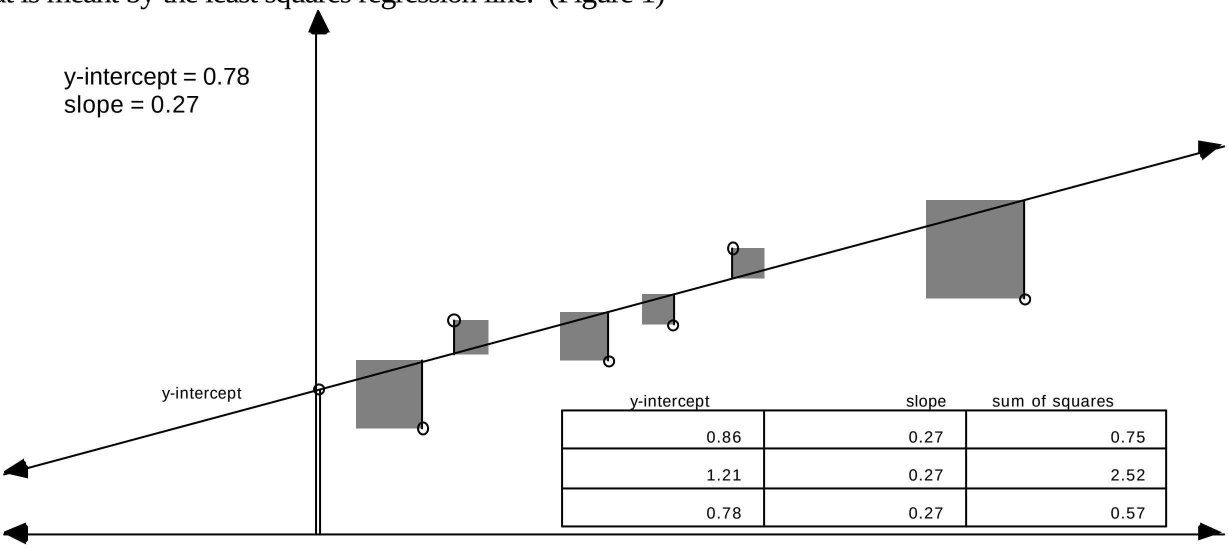
Technology

Despite its beginnings as a qualitative discipline, statistics is often characterized as "number crunching" – a field dependent on counting, ordering, and calculating. In fact, statisticians in the mid-1900's employed people as "calculators" whose work consisted of carrying out the many and involved calculations necessary for the application of statistical reasoning. Today, computers, calculators and graphing calculators provide access to computational power, spreadsheets, replay and recall of operations, ordering of data, organization and coding of data, changing scales, graphs, interaction between symbols and graphs, record keeping, motion. Each of these functions have made a difference in the nature of the statistics that it is possible to do and to teach and the way in which that content can be taught. Technology turns statistics into a hands-on activity based course where students become engaged in doing statistics not just learning about the formulas.

According to Rossman (1996), there are three basic uses for technology in statistics. Computers and graphing calculators allow students to

- Explore and investigate statistical phenomena. Students can explore several different graphical representations of a set of data or investigate different models for paired data. They can make predictions about a particular statistical property and then use a calculator to investigate their predictions, revising their predictions and iterating the process as necessary. Students can use calculators to investigate the best fitting lines, the effect of outliers, and the effect of sample sizes on confidence intervals. They can see for themselves whether dividing by n-1 is an unbiased estimator for the standard deviation of a sample by drawing many samples from a given set and checking the distribution of standard deviations. They can explore what happens if the conditions for using the test are not met; suppose $n\hat{p}$ is less than 10 (where n is the sample size, and $\hat{p}$ is the sample proportion) in finding a confidence interval for a proportion. What are the consequences?

- Carry out simulations. Technology allows students to conduct simulations which let them experience the long term behavior of sample statistics under repeated random sampling. Through simulation, students can build intuitions about probability and expected values – how likely is it to pass a true false test by guessing and how will the probability change as the number of questions increases. They can come to understand the behavior of sampling distributions and explore the patterns inherent in randomness.
- Visualize statistical concepts and perform calculations. Technology can produce graphical displays necessary to analyze real data sets, which are often large and use messy numbers. Using dynamic geometry software and the connection between graphs and symbolic expressions and their graphic representations, students can see how changing a regression line changes the sum of squared residuals – and can see visually what is meant by the least squares regression line. (Figure 1)



y-intercept	slope	sum of squares
0.86	0.27	0.75
1.21	0.27	2.52
0.78	0.27	0.57

Figure 1

Students can see how the shape of a distribution is related to the mean and standard deviation by changing the shape or the statistics and observing the effect. They can compare the t-distribution to the z-distribution and experiment with the degrees of freedom to see how the t-distribution changes. (Figures 2, 3)

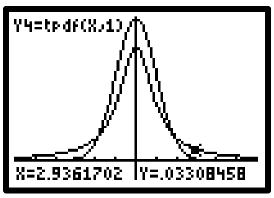


Figure 2

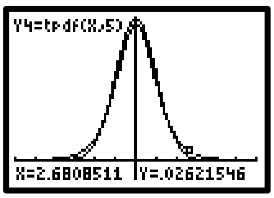


Figure 3

Computers and graphing calculators have had a direct impact on what we teach in introductory statistics. There are advantages to using either tool for instructional purposes. Despite the limitations of screen size, a graphing calculator is small enough and, comparatively speaking, inexpensive enough that every student can have access to one at all times. This is not yet the case with computers. What is the impact of technology on the curriculum?

Technology makes some of the things we used to do unnecessary.

Statistics books used to have detailed rules for finding the number of categories and the bin size for histograms. Creating histograms for messy data by hand takes a considerable amount of time. To maximize the likelihood that the histogram would reveal something about the data, certain structures were suggested for generating the plot that added a degree of complexity to the process. Today, with a keystroke, students in lower secondary school using either a graphing calculator or a computer, can investigate what different bin sizes do to the distribution and in the process come to more fully understand the messages inherent in the data.

Consider the formulas once taught as easier ways to carry out complex calculations. The use of some of these formulas probably contributed to the lack of understanding of the statistic being produced (and reinforced perceptions on the part of many students that statistics was a difficult subject to be avoided if possible). Take for example, the standard deviation. The definition is simple and makes quite good sense, when posed as a way to find a measure of the variation from the mean. Find the difference from the mean; the differences sum to zero, squaring the differences eliminates the negative sign; divide by n to even out the differences among the data points, and finally, take the square root to return to the original scale. Now, consider the alternate formula, described as a practical way to simplify the calculations (Willoughby, 1968).

$$\text{standard deviation} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \quad \text{or} \quad \frac{1}{n}\sqrt{\sum \ x_i^{\,2} - \bar{x}^{\,2}}$$

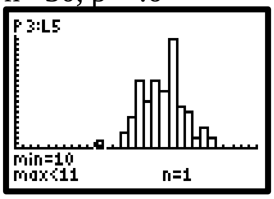
Mathematically equivalent, the two procedures send very different messages about the underlying concept. Unfortunately, many students never had the opportunity to work with the formula based on the definition. Today, spreadsheets allow students to explore the definition in a variety of cases, look at the magnitude of the differences squared and their contribution to the squared sum, see the effect of outliers, and in general, build an understanding of exactly what standard deviation represents.

Technology allows us to do old things in new ways.

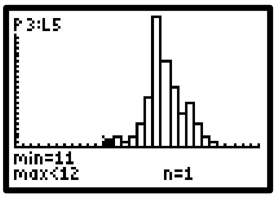
Technology allows students to quickly investigate different plots of the same data, moving from histograms to box plots. It allows students to compare data sets by simultaneously looking at multiple box plots. Technology allows students to use large data sets and to access real information from the web. The concept of the mean as the balance point or centroid of a distribution comes to life with technology. Technology makes it critical that students learn to read statistical analysis output from a statistical package such as Data Desk or Minitab.

Technology allows students to simulate sampling distributions to see how likely a given outcome will be. For example, suppose that 60% of the cars produced by automobile manufacturers are red. Would finding only 14 red cars in a random sample of 30 cars be a likely event?

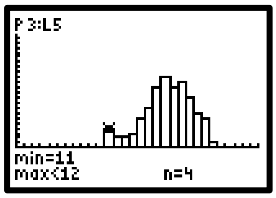
n= 30, p = .6



90% of the samples
have 14 to 23 red cars
Figure 4



90% of the samples
have 14 to 22 red cars
Figure 5



90% of the samples
have 12 to 23 red cars
Figure 6

Figures 4, 5, and 6 show three sets of one hundred simulations of 30 cars selected at random from a population where 0.6 are red. In each case, 14 red cars seems likely to occur with some level of regularity. To have only 9 cars seems quite unlikely based on the evidence in the simulations, giving students an introduction to outcomes that can be considered rare events and a sense of the relatively arbitrary designation of what is "likely". If something that occurs at least 90% of the time is defined as likely, a box plot for the likely outcomes can be constructed and used as a way to describe the level of acceptable variation around the expected outcome of 18 red cars. This approach can be used to provide the infrastructure for defining confidence intervals in a very concrete and functional way. (Landwehr et al, 1987)

Simulation can also quickly and easily generate sampling distributions for sample means and enable students to develop an understanding of the Central Limit Theorem. It provides a visual and concrete demonstration of how, regardless of the shape of the original distribution, the distribution of sample means will be a normal distribution. The population described in the histograms below (Figures 7, 8) are the areas of 100 random rectangles ranging in size from 1 to 18 square units. (Scheaffer, 1997).

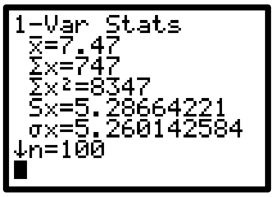


Figure 7

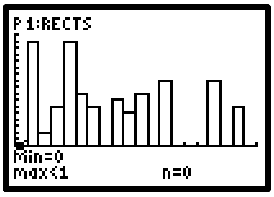


Figure 8

If the algorithm in Figure 9 is preceded by initializing the counter A to be 1, it will generate random samples of size 15 from the population of rectangles, calculate the mean area of each sample, and record the mean in a list. The sample size can be changed by changing the 15, allowing students to experiment with different size samples, building intuition about the shape, center, and spread of the distributions. (Figures 10, 11, 12)

N=5

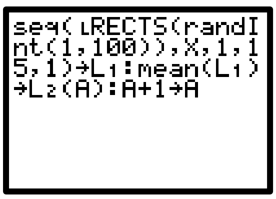


Figure 9

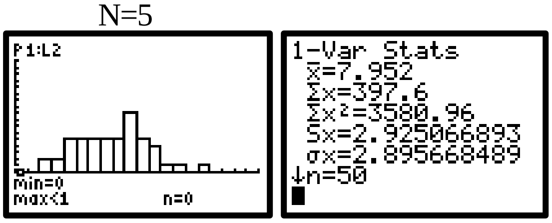


Figure 10

N = 10 N = 15

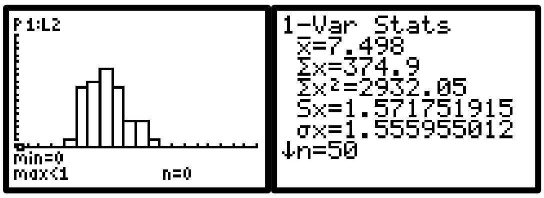


Figure 11

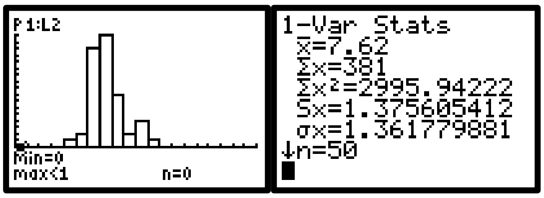


Figure 12

Students quickly notice that the mean in each case is near the population mean, while the standard deviation decreases as the sample size increases. (With some perseverance and collaboration, students can actually use regression to find the relation between sample size and standard deviation.)

Just as scientific calculators made logarithmic tables unnecessary, graphing calculators make most statistical tables unnecessary. In addition, graphing calculators can both graph and calculate the probabilities for a given distribution reinforcing the conceptual understanding of what the numbers actually represent (Figures 13, 14).

Normal Distribution, $\mu = 0$, $\sigma = 1$ Normal Distribution $\mu = 50$, $\sigma = 5$

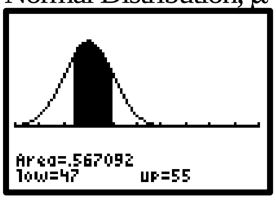


Figure 13

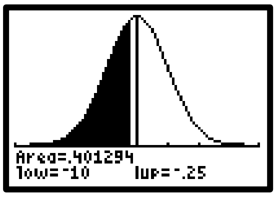


Figure 14

These capabilities allow a shift in emphasis in the questions typically asked on assessments from understanding how to read the table to understanding what the distribution reflects. They remove routine procedures and provide opportunities for interpretation and problem solving. This means, however, that the very nature of the questions must be different. "For n = 25, p = 0.4, find $P(x \leq 21)$" has to be replaced with questions such as the following. For a normally distributed variable with a mean of 50 and a standard deviation of 5, will the probability of an outcome from 45 to 55 be more likely than that of a normally distributed random variable with the same mean and a standard deviation of 10? What do you know about a normal distribution that will explain why your answer makes sense? Compare the probability to the probability of an outcome from -1 to 1 for a normally distributed random variable with mean 0 and standard deviation 1.

Technology allows us to do new things we were unable to do without it.

The use of the median as a measure of center was only possible with the appearance of technology that could sort and count. It seems plausible that one of the reasons why much of the current statistical theory is mean based is that until very recently, the mean was the only useful measure of center it was possible to calculate in an efficient and practical way. There are clearly opportunities to develop new statistical theories, particularly in light of the ability of technology to perform high powered numerical calculations and analysis. Regression equations are now part of the secondary curriculum because technology makes it possible for students to create mathematical models quickly and easily (Hopfensperger et al, 1999). They can describe the relationship between variables and assess the model by calculating residuals and looking at their graph. Consider, for example the relation between the number of spaces from

Go and the price of a property in the game of Monopoly. A linear regression yields Price = 6.78d+67, where d is the number of spaces from Go, with a correlation coefficient of 0.88.

Monopoly

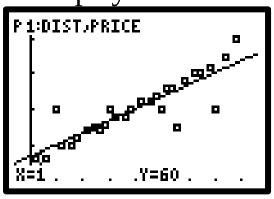


Figure 15

Monopoly

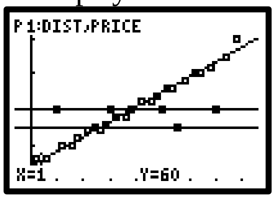


Figure 16

Looking at the plot (Figure 15) and reflecting on the nature of the data, however, brings out the fact that railroads uniformly cost $200 and the utilities are $75. Thus, there are really three relationships involved, two of which are constant. A better model would be to use separate functions for railroads and utilities and find Price = 8.4d+39.8 with r = .99 for the other properties. (Figure 16) Because calculators can produce logarithms and re-scale data, all students can also investigate the underlying principles of modeling non linear relations, studying exponential, logarithmic, and power models.

Technology allows students to investigate the normal curve

$$y = \frac{1}{\sqrt{2p\,s}} e^{-(x-m)^2/2s^2} .$$

Students can understand and carry out multiple regression techniques using matrices to organize and manage information and technology to perform the computations. Consider the problem of trying to predict college grade point averages using both the verbal and math scores on the SAT. Table 1 contains the data on GPA, SATV, and SATM for each of 15 college students (Witmer et al., 1998).

Table 1: College Grade Point Average

Student number	GPA	SATV	SATM
1x	3.58	670	710
2	3.17	630	610
3	2.31	490	510
4	3.16	760	580
5	3.39	450	510
6	3.85	600	720
7	2.55	490	560
8	2.69	570	620
9	3.19	620	640
10	3.50	640	660
11	2.92	730	780
12	3.85	800	630
13	3.11	640	730
14	2.99	680	630
15	3.08	510	610

Source: Oberlin College, 1993

The problem is to find some b_0, b_1 and b_2 such that

$b_0(1) + b_1V + b_2M = \hat{GPA}$

This can be written in matrix form as

$$\begin{bmatrix}1, V, M\end{bmatrix} \cdot \begin{bmatrix}b_0, b_1, b_2\end{bmatrix} = \begin{bmatrix}\hat{GPA}\end{bmatrix}$$

This can be expressed as AX = B where A is the matrix determined by using a constant 1 and columns of verbal and math scores, X the coefficient matrix, and B the predicted grade point averages, . Using matrix properties,

$$\begin{aligned} AX &= B \\ A^tAX &= A^tB \\ (A^tA)X &= A^tB \\ (A^tA)^{-1}(A^tA)X &= (A^tA)^{-1}A^tB \\ X &= (A^tA)^{-1}A^tB \end{aligned}$$

Solving for matrix X, using the GPA data in matrix form and the formula gives

$\mathbf{b} = \begin{bmatrix}1.45112, 0.00190, 0.00083\end{bmatrix}$

and the desired regression equation for predicting college grade point as a function of math and verbal SAT scores is

$$\hat{GPA} = 1.45112 + 0.001906V + 0.00083M.$$

(An interesting exercise is to compare this model to the regression of GPA Vs total SAT score.) Matrix A can be any dimension, so A could include other relevant data as well. Thus, for a given set of data, a regression model can be quickly and easily computed - providing the matrix system has a solution.

Conclusion

Clearly technology makes a difference in statistics education. It opens the doors to statistics and statistical reasoning to all students, provides new insights into working with data and makes it possible to use real data and situations of interest to motivate both the study of statistics and of mathematics. But using technology is not without its challenges. There is a question of equity; how do we ensure that all students and teachers have the same access to the power of technology so they have the same opportunities? There is a question of integrity; is it fair to allow technology to be used at all times, no matter what the task, on routine work and in assessment situations? Should there be a limit on the kind of technology students can use in certain situations? There is the question of inappropriate use. In some cases, statistical reasoning and the opportunity for investigation will be replaced by button pushing, reducing the study of statistics to another set of mysterious steps that are now calculator routines. There is the question of how to track student work and thinking. What kind of a paper trail will inform a teacher about student thinking? There is the question of students who do very foolish things relying on technology seemingly in lieu of thinking. What would their excuse have been before technology was on the scene? Would these students even be here if it were not for technology? There is the question of continually advancing technology. How do teachers stay current? How does a curriculum acknowledge and capitalize on the advances of technology without being in a constant state of change?

Statistics has evolved over the ages from being the study of the characteristics of a state taught in law schools to political economics, to the philosophy of social relations, and eventually strongly linked to mathematics. Today, technology is again pushing the bounds of the discipline as it is currently known - the one clear message is that despite all of the many challenges, yesterday's content with yesterday's tools cannot prepare today's students for tomorrow's world.

References

Burrill, G. (Ed.) (1991). *Guidelines for the teaching of statistics*. Palo Alto, Ca.: Dale Seymour Publications.

Burrill, G., Hopfensperger, P., & Landwehr, J. (1999). *Exploring least squares regression* . Palo Alto, CA.: Dale Seymour Publications.

College Entrance Examination Board and Educational Testing Service (1996). *Advanced Placement course description statistics May 1997, May 1998*. Princeton, NJ.: The College Board.

Fathom. (In development.) Emeryville CA.: Key Curriculum Press. jbrinker@keypress.com

Freedman, David, Robert Pisani, and Roger Purves. *Statistics*. New York, NY.: W.W. Norton & Company, Inc., 1980.

Landwehr, J., Swift , J., & Watkins, A. (1987). *Exploring samples and information from surveys.* Palo Alto, CA.: Dale Seymour Publications.

Moore, D.S. (1997). *The active practice of statistics.* New York NY: W.H.Freeman and Company.

Moore, D.S. & McCabe, G.P. (1993). *Introduction to the practice of statistics*. New York NY: W.H.Freeman and Company.

Ottaviani, M.G. "A history of the teaching of statistics in higher education in Europe and the United States 1660 to 1915." (1989). In Morris, R. (Ed.) *Studies in mathematics education: the teaching of statistics*. Vol 7. Paris, Fr.: UNESCO

Rossman, A. (1996). *Workshop statistics*. New York, NY.: Springer-Verlag.

Scheaffer, R. L., Gnanadesikan, M., Watkins, A., & Witmer, J.A. (1996). *Activity-based statistics*. New York, NY.: Springer-Verlag.

Wallechinsky, D. & Wallace, I. (1975). *The people's almanac.* Garden City, NY.: Doubleday & Company, Inc. p.79.

Witmer, J., Burrill, G., Burrill, J., & Landwehr, J. (1998). *Advanced modeling using matrices.* Palo Alto, CA.: Dale Seymour Publications.