An On-Line Virtual Environment For Teaching Statistical Sampling And Analysis

Michael T. Marsh, Shippensburg University, USA

ABSTRACT

Regardless of the related discipline, students in statistics courses invariably have difficulty understanding the connection between the numerical values calculated for end-of-the-chapter exercises and their usefulness in decision making. This disconnect is, in part, due to the lack of time and opportunity to actually design the experiments and collect the data. The prototypes proposed in this project were developed to allow students to design experiments and collect data in relevant settings without the impediments to real data collection. The virtual environments attempt to replicate real situations of interest in which students can design and run experiments, devise alternative sampling strategies, analyze the results of experiments, and relate the result to the original experiment. The setting and underlying data set detailed in this paper were developed to allow students to experience a wide range of statistical concepts typically found in introductory statistic courses, such as basic descriptive statistics, estimation, hypothesis testing, ANOVA, and regression. Assessments of student knowledge after using this approach have shown marked increases in students' understanding of statistical concepts, especially confidence intervals and hypothesis testing. Specific details about the data set are provided as are suggestions for using it in an introductory statistics class. Potential uses and examples for a variety of disciplines are also included.

Keywords: virtual, online, dataset, introductory statistics, statistical inference

INTRODUCTION

Many articles and on-line resources reward instructors with techniques adaptable to teaching introductory statistics. The hands-on, activity-based ideas coupled with the range of interactive activities available on the internet have enhanced student motivation and understanding. (delMas, Garfield, & Chance, 1999) Lock (2001) directs browsers/readers to websites that typify the various sorts of online activities which are currently available via the World Wide Web to help support statistics instruction. The majority of research notes that certain fundamental statistical concepts, especially sampling distribution, confidence levels, and hypothesis testing, are notoriously difficult for students to truly comprehend at an intuitive level. (Gourgey, 2000) (Anderson-Cook, 1999) and instructors are continuously exploring innovative teaching practices in the interest of rectifying this situation. In particular, a number of authors have reported success in engaging students in simulation exercises, designed to convey the concepts of sampling distributions and sampling variability.

A review of student knowledge levels in introductory statistics courses was adequate when asked to solve problems at the end of a chapter in typical statistics textbooks. However, when asked to accomplish activity-based statistics similar to the exercises proposed by Smith (1998), many were at a total loss about how to start. As noted by several researchers, the transition from textbook to practice is difficult for many students. The importance of designing experiments, gathering data for analysis, obtaining results, and relating the results to the original experiment is highlighted by Schwarz (2003). Schwarz (2007) again emphasized the importance of, but inherent problems with, gathering and using real data. Successful experiences with providing each student in a large, multi-section class with a unique dataset for homework and in-class exercises have been shown by Vaughan (2003). Also, some of the learning problems that typically arise when presenting difficult concept can be mitigated by using settings familiar to students. Martin (2003)

THE SETTING

StatVillage (Schwarz, 1997) has proven to be an excellent resource for combining sampling and subsequent analysis at more advanced levels and the idea should prove useful for teaching more elementary concepts. The resources and associated activities presented in this paper focus on sampling techniques and statistical analysis using a unique, customized dataset. In a hypothetical college campus named StatU the setting and variables readily familiar to students. The online virtual "campus" and associated dataset of StatU was created to meet the apparent need for an intermediate step between end-of-the-chapter problems and real world experiment design and data gathering. StatU emulates the finest traditions of StatVillage, but adapts the concepts to meet the needs of a more introductory level and be used as a common thread throughout a complete introductory statistics course encompassing a wide range of statistical concepts.

A university setting was chosen as it provides an environment and terminology familiar to all students in a class. StatU simulates four dormitories on the quad. Each dorm has 10 floors with 12 rooms to a floor. Two students live in each room making a student population of 960. Each student at StatU has ten attributes; four can be used as identifiers in sampling strategies:

1. Dorm (North, South, East, West)
2. Floor (1 – 10)
3. Room (A – L)
4. Bunk (R, L)

and six can be used as variables for statistical analysis:

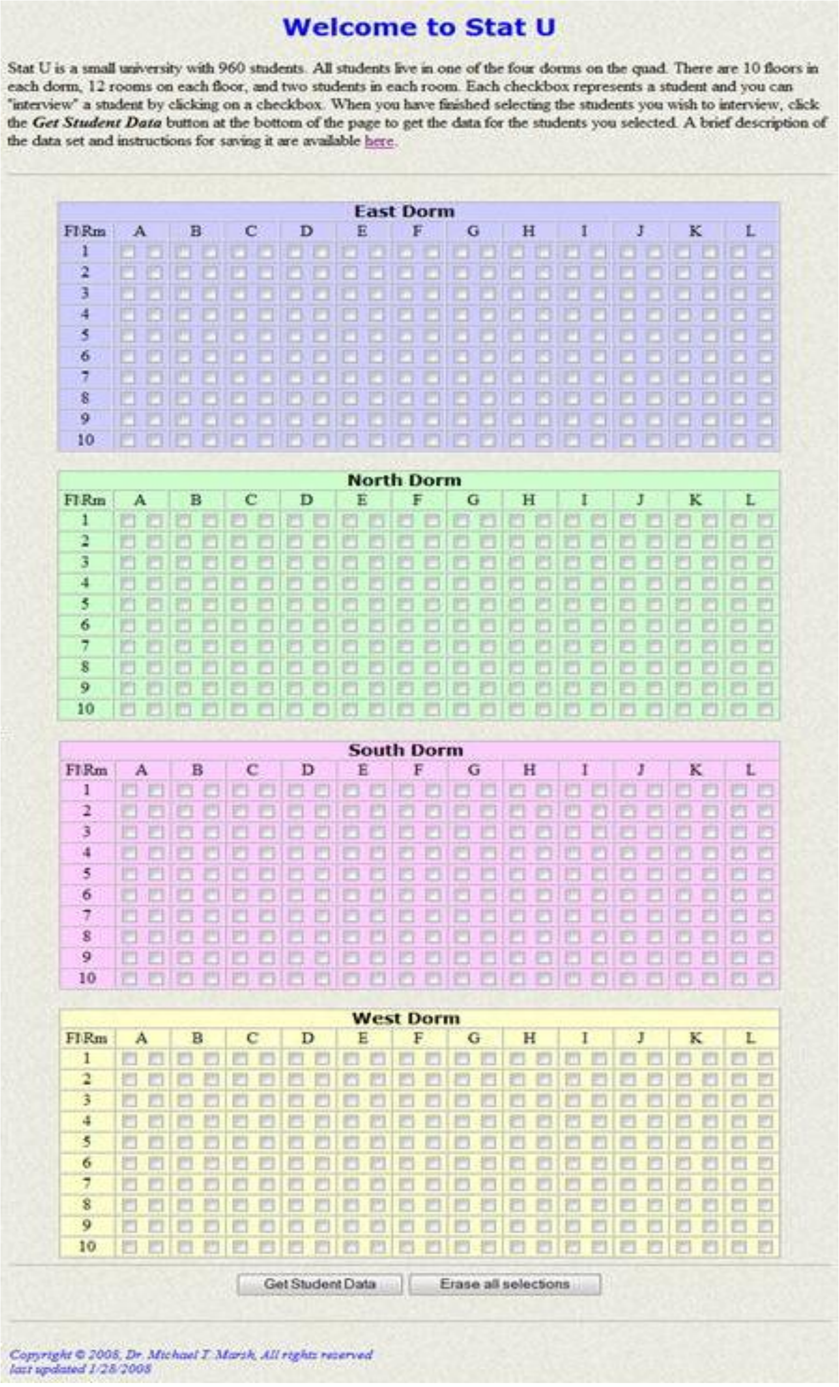
5. Family household income
6. SAT score
7. GPA
8. Number of siblings
9. Number of study hours study per week
10. Binge drinker

THE DATA SET

While the data is hypothetical, several parameters associated with the attributes reflect real world values. The family household income distribution was derived from information found at http://www.census.gov/hhes/www/income/income.html. The SAT information can found at www.collegeboard.com, as can be the relationship between SAT scores and income. The GPA data are based on factual GPA data at a Pennsylvania college and the correlation between GPA and SAT scores in the dataset reflects that stated by the Educational Testing Services for a select subset of universities. The distribution of siblings reflects census data for family size and income available at the federal government census site. Binge drinking data is similar to that reported in a recent Harvard study (Wechsler, 2000). The correlation values, such as between grades and amount of study or between grades and binge drinking, are hypothetical but loosely based on anecdotal evidence. The types of attributes were selected to allow presenting concepts that include discrete, continuous, or binomial variables.

USING STATU IN THE CLASSROOM

In a typical introductory statistics course, sampling may be only a small part of the curriculum and discussed only briefly. Like StatVillage, StatU provides a setting that easily enables sampling using various types of sampling methodologies. The instructor may assign an analysis of just those students with GPAs greater that 3.00 to analyze Study Hours in a stratified sample, for example. Cluster sampling can be accomplished by designating one dorm or specified floor(s) in dorms. The online layout enables both random and systematic sampling to be accomplished quickly and easily with little instructor guidance. Having the population already partitioned, particularly by dorms, streamlines the development of experiments that compare two or more statistics.



Welcome to Stat U

Stat U is a small university with 960 students. All students live in one of the four dorms on the quad. There are 10 floors in each dorm, 12 rooms on each floor, and two students in each room. Each checkbox represents a student and you can "interview" a student by clicking on a checkbox. When you have finished selecting the students you wish to interview, click the ***Get Student Data*** button at the bottom of the page to get the data for the students you selected. A brief description of the data set and instructions for saving it are available here.

East Dorm

Fl\Rm	A	B	C	D	E	F	G	H	I	J	K	L
1	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
2	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
3	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
4	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
5	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
6	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
7	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
8	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
9	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
10	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □

North Dorm

Fl\Rm	A	B	C	D	E	F	G	H	I	J	K	L
1	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
2	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
3	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
4	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
5	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
6	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
7	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
8	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
9	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
10	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □

South Dorm

Fl\Rm	A	B	C	D	E	F	G	H	I	J	K	L
1	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
2	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
3	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
4	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
5	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
6	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
7	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
8	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
9	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
10	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □

West Dorm

Fl\Rm	A	B	C	D	E	F	G	H	I	J	K	L
1	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
2	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
3	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
4	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
5	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
6	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
7	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
8	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
9	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □
10	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □	□ □

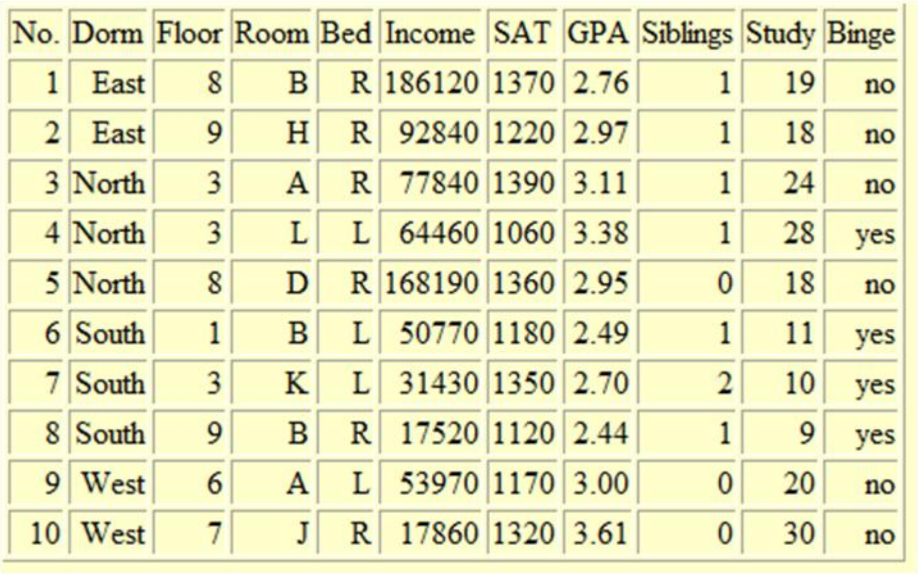
Copyright © 2008, Dr. Michael T. Marsh, All rights reserved
last updated 1/28/2008

Figure 1

A view of the online representation of the StatU campus is shown in Figure 1 and the interactive environment can be accessed at http://webspace.ship.edu/mtmars/StatU/StatU.html. The online description and instructions for using Stat U are as follows: *When a student is selected, i.e. "interviewed", (s)he will truthfully answer questions about household income (Income), SAT score (SAT), current grade point average (GPA), number of brothers and sisters (Siblings), the number of hours (s)he studies per week (Study), and whether or not (s)he has been binge drinking during the past two weeks (Binge). (Note: A recent Harvard School of Public Health College Alcohol Study defines binge drinking as five drinks in a row for males, four for females.) While the attributes associated with each student are fictional, several distributions, averages, and correlations are based factual data. For additional information concerning the development of the dataset, contact Dr. Michael T. Marsh.*

To analyze the sample data, select and copy the table with the observation values, then paste it into an Excel worksheet.

Most statistics courses introduce descriptive statistics early in the course. Students are usually able to solve end-of-the-chapter exercises quite easily to find basic descriptive measures such as means, ranges, or standard deviations. Frequently, however, it is another matter when asked to work with real data. StatU has proven to be an effective transition to more realistic problems as well as an opportunity to find values using spreadsheets such as Microsoft® Excel or statistics analysis software such as SPSS or SAS. Another benefit is that as students find various values for their individual samples, the ground work for concepts related to sampling distributions is formed. Figure 2 shows the StatU output for a random sample of size 10. Figure 3 shows the results with the data copy/pasted into an Excel worksheet and the Data Analysis – Summary Descriptive Statistics applied to GPA. Of course any of the Excel functions such as =AVERAGE or =STDEV can easily be applied to a selected attribute.



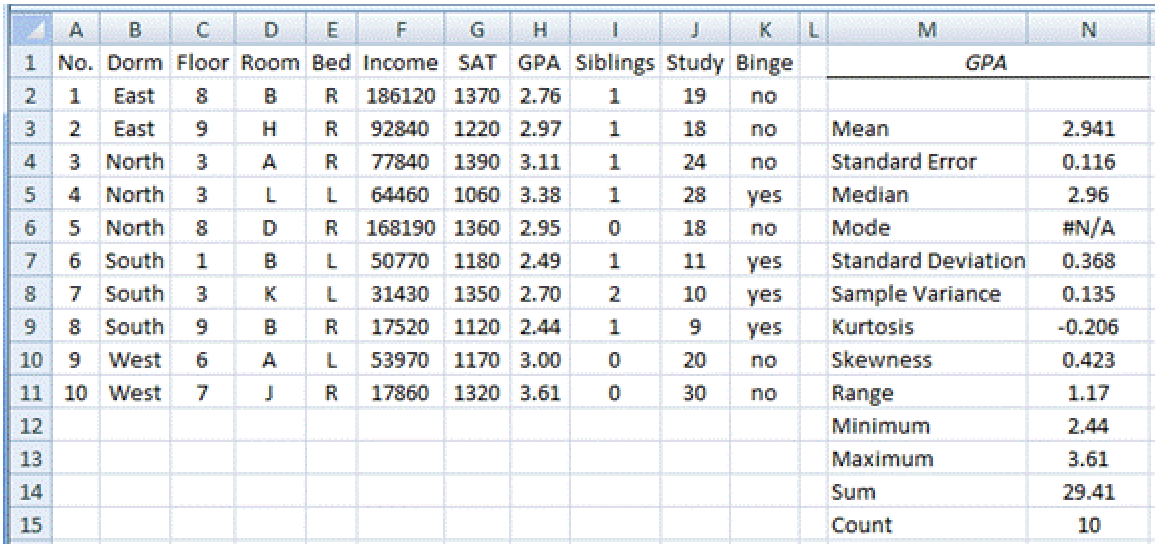
No.	Dorm	Floor	Room	Bed	Income	SAT	GPA	Siblings	Study	Binge
1	East	8	B	R	186120	1370	2.76	1	19	no
2	East	9	H	R	92840	1220	2.97	1	18	no
3	North	3	A	R	77840	1390	3.11	1	24	no
4	North	3	L	L	64460	1060	3.38	1	28	yes
5	North	8	D	R	168190	1360	2.95	0	18	no
6	South	1	B	L	50770	1180	2.49	1	11	yes
7	South	3	K	L	31430	1350	2.70	2	10	yes
8	South	9	B	R	17520	1120	2.44	1	9	yes
9	West	6	A	L	53970	1170	3.00	0	20	no
10	West	7	J	R	17860	1320	3.61	0	30	no

Figure 2

The variability among sample data is frequently not apparent if traditional approaches are used, with all students using the same data. Even when using StatU students are not able to see and compare their results with those of other students. An inexpensive and effective way to view all students' results is to have each student input their results into the same spreadsheet and compare them with other students using Google Documents. Google Documents is a free software application, available at http://www.google.com that allows simultaneous inputs into a spreadsheet.

Another feature of the StatU dataset is the variety of population distributions available providing opportunities for experiments with sampling distribution concepts or more sophisticated tests for goodness-of-fit.

Figures 4, 5, and 6 show examples of normal, highly skewed, and slightly skewed distributions that are inherent in the StatU dataset.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	No.	Dorm	Floor	Room	Bed	Income	SAT	GPA	Siblings	Study	Binge		*GPA*	
2	1	East	8	B	R	186120	1370	2.76	1	19	no			
3	2	East	9	H	R	92840	1220	2.97	1	18	no		Mean	2.941
4	3	North	3	A	R	77840	1390	3.11	1	24	no		Standard Error	0.116
5	4	North	3	L	L	64460	1060	3.38	1	28	yes		Median	2.96
6	5	North	8	D	R	168190	1360	2.95	0	18	no		Mode	#N/A
7	6	South	1	B	L	50770	1180	2.49	1	11	yes		Standard Deviation	0.368
8	7	South	3	K	L	31430	1350	2.70	2	10	yes		Sample Variance	0.135
9	8	South	9	B	R	17520	1120	2.44	1	9	yes		Kurtosis	-0.206
10	9	West	6	A	L	53970	1170	3.00	0	20	no		Skewness	0.423
11	10	West	7	J	R	17860	1320	3.61	0	30	no		Range	1.17
12													Minimum	2.44
13													Maximum	3.61
14													Sum	29.41
15													Count	10

Figure 3

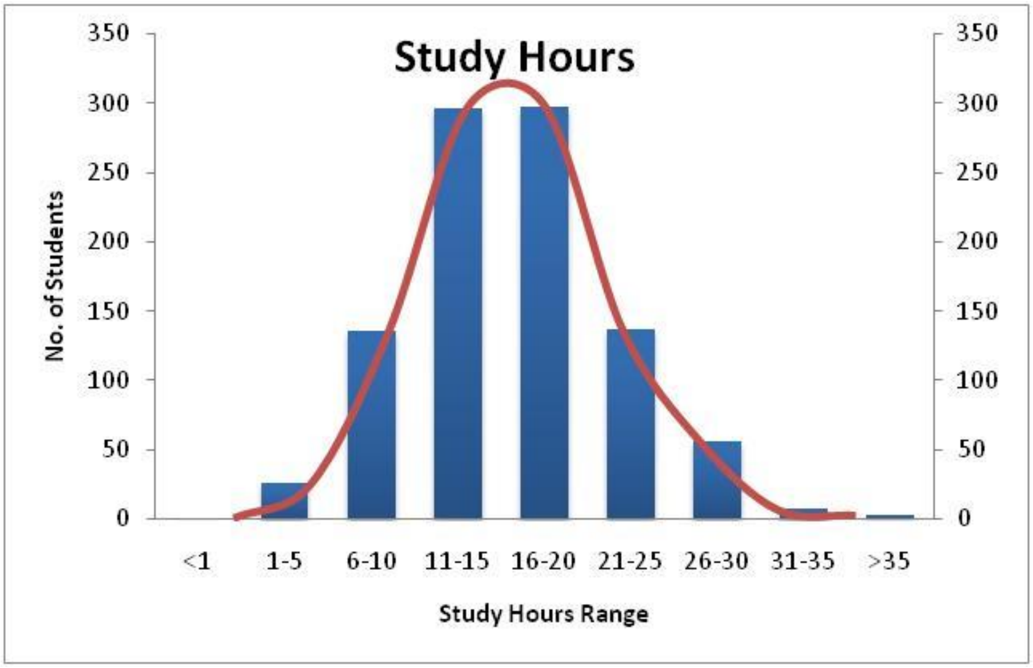


Figure 4

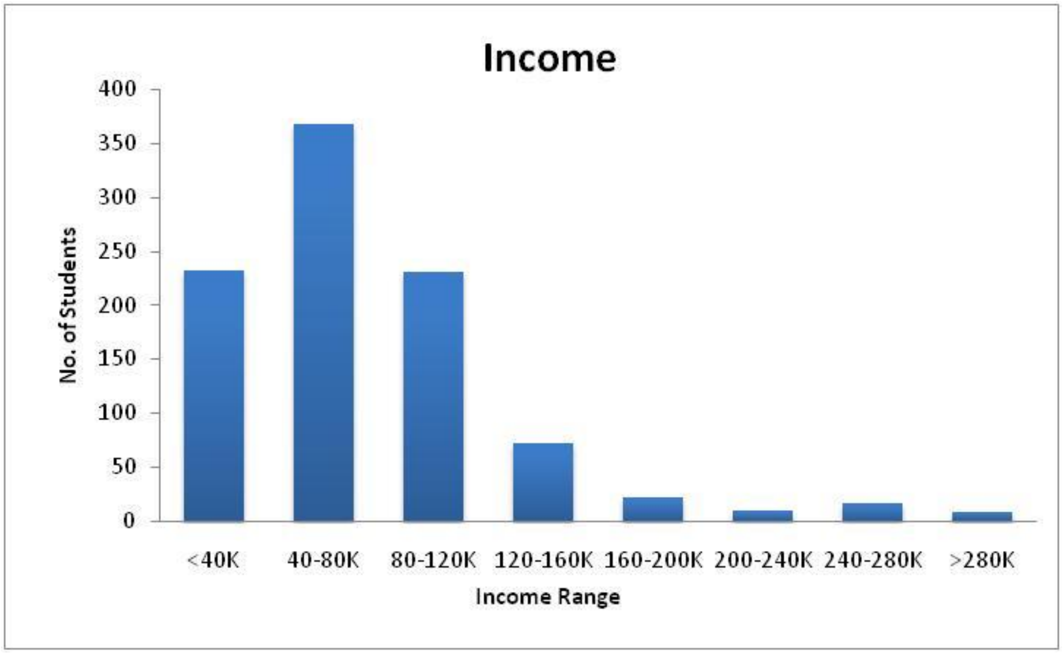


Figure 5

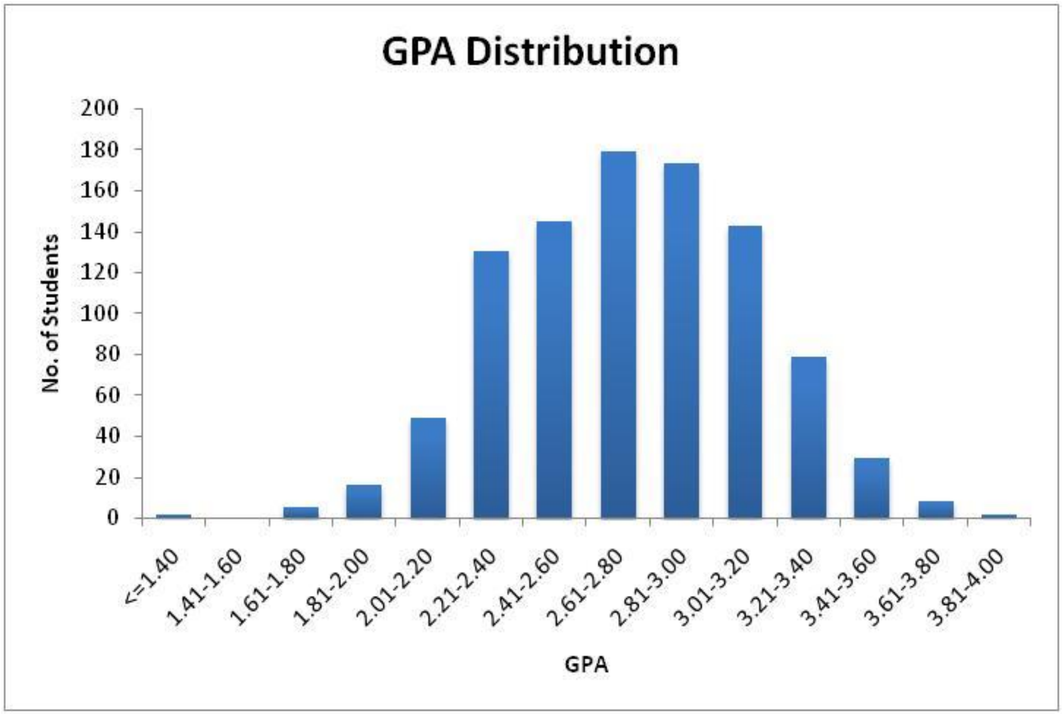


Figure 6

The learning payoff from StatU comes when teaching inferential statistics. The concept of sampling distributions and related confidence intervals and hypothesis testing are among the most difficult for most students to grasp. The implications of the variation in values that result with confidence intervals and hypothesis testing analysis are the concepts with which StatU has been most effectively used. For example, once student are comfortable with the computations involved in finding the upper and lower limits for confidence intervals, they are then instructed to take a sample, usually random, of a specified size using StatU and compute a confidence interval with specified alpha. Once students calculate the upper and lower limits of the confidence interval for their individual samples, the intervals are input and display together on a spreadsheet. Again, Google Document provides an efficient way for students and the instructor to access a spreadsheet simultaneously. The instructor can open a spreadsheet with students' names in say Column A of the spreadsheet and have students input their calculated data. In a class of 30 students it is typical that at least one confidence interval does not include the population mean. Since the instructor knows the population mean it can easily be shown how not all confidence intervals contain the population mean, thus helping students understand the concept.

Similarly, students' understanding of the concept involved in hypothesis testing has been increased also. The StatU environment lends itself to all variations of confidence intervals and hypothesis testing including:

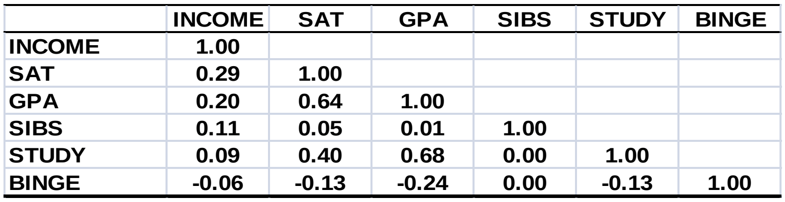
- Normal or non-normal distributions
- Testing population means or variances
- Population variance known or unknown (the instructor has the population parameters)
- z or t with varying sample sizes
- Population proportion with the binge drinking binomial variable
- Differences between two population parameters such as between dorms or floors

Some of the analysis can be made even more meaningful if the instructor associates characteristics of the dataset with the students' environment, such as declaring the North dorm to be a "Quiet" dorm.

Other concepts that use StatU effectively are:

- ANOVA, among dorms or floors, for example.
- Correlation
- Simple regression analysis
- Multiple regression analysis

Figure 7 is the correlation matrix for the dataset



	INCOME	SAT	GPA	SIBS	STUDY	BINGE
INCOME	1.00					
SAT	0.29	1.00				
GPA	0.20	0.64	1.00			
SIBS	0.11	0.05	0.01	1.00		
STUDY	0.09	0.40	0.68	0.00	1.00	
BINGE	-0.06	-0.13	-0.24	0.00	-0.13	1.00

Figure 7

The StatU dataset has been specially constructed to provide examples of many different statistical situations. For example, the population means of GPA for the North and West dorms are not statistically differently, albeit just barely, at alpha = 5%. (Table 1) Typically, however, several student hypotheses that the difference is zero will be rejected. (Table 2)

z-Test: GPA population means		
	West	*North*
Mean	2.716	2.773
Known Variance	0.143	0.15
Observations	240	240
Hypothesized Mean Difference	0	
z	-1.634	
P(Z<=z) one-tail	0.051	
z Critical one-tail	1.645	
P(Z<=z) two-tail	0.102	
z Critical two-tail	1.960	

Table 1

z-Test: GPA sample means		
	West	*North*
Mean	2.6328	2.8792
Known Variance	0.143	0.155
Observations	25	25
Hypothesized Mean Difference	0	
z	-2.257	
P(Z<=z) one-tail	0.012	
z Critical one-tail	1.645	
P(Z<=z) two-tail	0.024	
z Critical two-tail	1.960	

Table 2

Simple and multiple regression experiments are also straightforward to conduct using StatU and Excel. Figure 8 shows a typical graphical result for the traditional GPA - SAT populations.

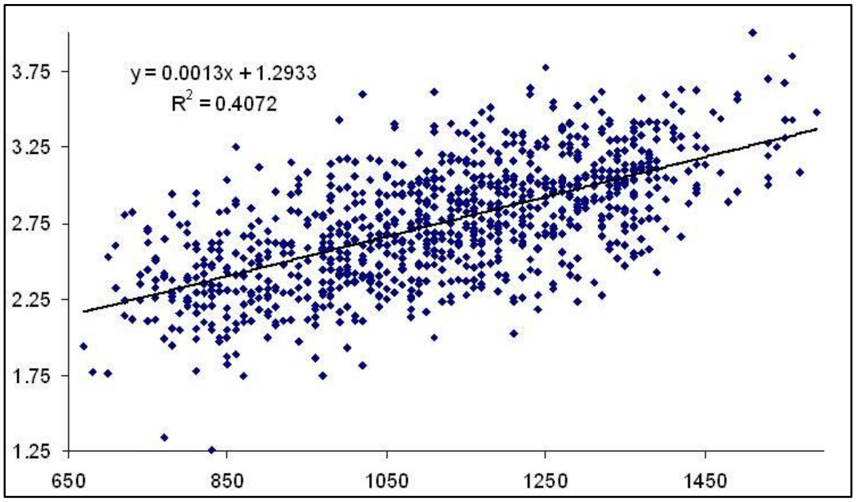


Figure 8

Figure 9 shows the multiple regression values of explanatory variables study hours, binge drinking, and SAT scores with GPA as the dependent variable for the StatU population.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.774368094					
R Square	0.599645945					
Adjusted R Square	0.513855791					
Standard Error	0.216127993					
Observations	18					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	0.9795	0.3265	6.9897	0.0042	
Residual	14	0.6540	0.0467			
Total	17	1.6335				
Parameter	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.6562	0.4984	3.3231	0.0050	0.5872	2.7251
SAT	0.0007	0.0004	1.7052	0.1102	-0.0002	0.0016
Study	0.0214	0.0092	2.3354	0.0349	0.0017	0.0411
Binge	-0.1015	0.1259	-0.8060	0.4337	-0.3715	0.1686

Figure 9

The StatU dataset has specific characteristics to demonstrate the concept of multicollinearity. The pattern of standardized residuals in Figure 10 indicates the almost certain correlation among the error terms for the sample.

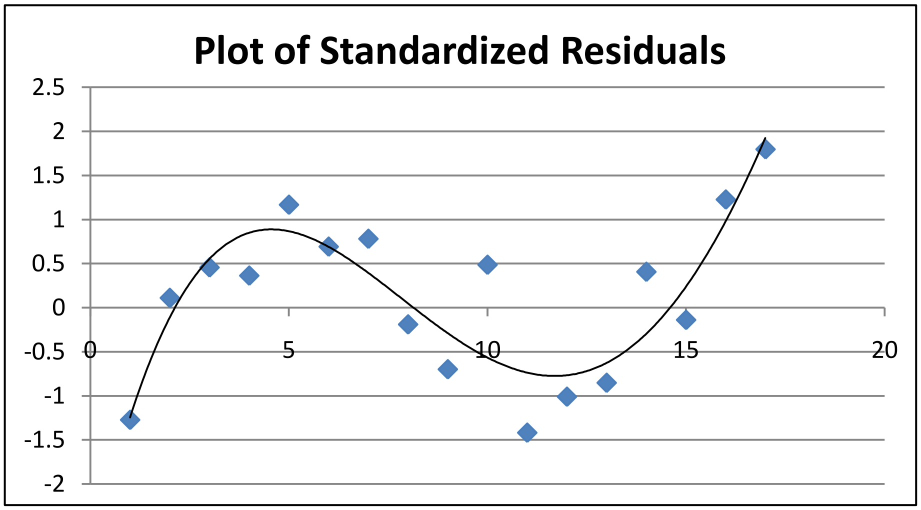


Figure 10

The dataset also includes examples of more esoteric statistical concepts. Figure 11 demonstrates the concept of heteroskedasticity, for example.

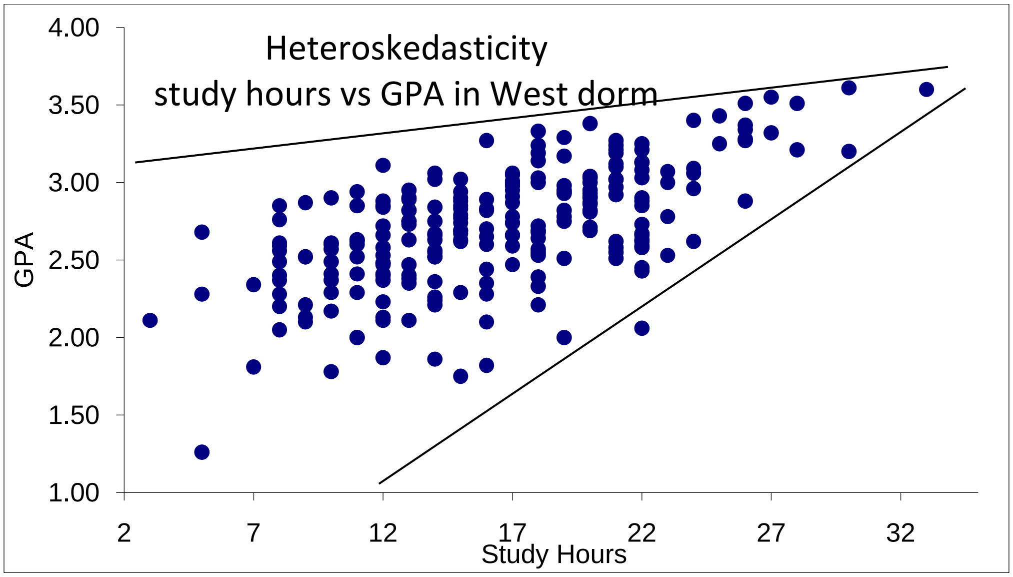


Figure 11

EXTENSIONS TO OTHER SETTINGS

While the StatU setting was developed for introductory statistics courses in business, the concept could be used in any course where statistical analysis is integral to the curriculum. Figure 12 is an example where the sampling grid overlays a lake system, a setting that could be used in a biology course where students sample waterfowl nesting sites. In a hydrology course, the background could be a watershed or for a marketing course, the grid could overlay a cityscape. Other examples would be a landform background for soil analysis, an archeological site for an anthropology course, or geographical areas for political polling. Existing datasets could be associated with the sampling grid.

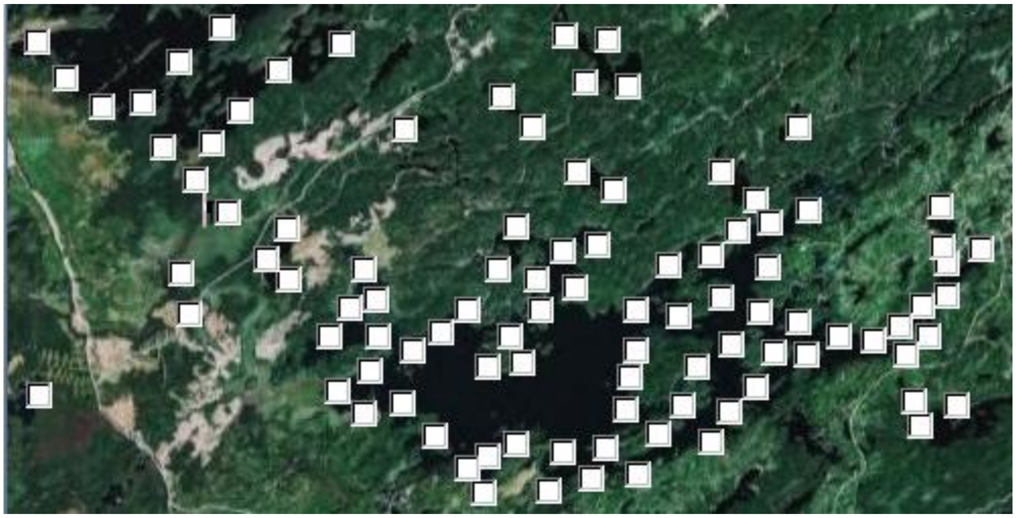


Figure 12

CONCLUSION

StatU provides several advantages for teaching introductory statistics that are not yet available in other online activities. The greatest single learning benefit is the improvement in the understanding of statistical inference but perhaps more importantly, class participation has increased, anxiety levels appear to be lower, and students show genuine curiosity about the concepts related to their analysis. Students are comfortable working with data having variables they can relate to. That the dataset values reflect real-world parameters yet do not include those real-world data collections problems such as missing data, extreme outliers, erroneously recorded data, non-responses, etc is appropriated for a first course in statistics. Students are able to obtain results of individual samples in a very short time period and the variability among different samples from the same population becomes more evident. StatU offers many of the desirable features listed by Darius (2007); it is available wherever the internet is available, many different data collection designs are allowed, the instructor has complete knowledge of the population and can assign experiments to emphasize specific concepts, different results are generated for each student or for the same student with each replication, and the virtual environment is general enough to allow a wide range of activities. Students appreciate the ability to be actively involved in the learning activity and many times take additional samples "just to see what happens". After the course, sometimes years later, students comment on how much they enjoyed using StatU and how statistics is not as difficult as they had thought it would be.

AUTHOR'S NOTE:

Contact Dr. Marsh (email:mtmars@ship.edu) if you would like the entire dataset emailed to you as an Excel file.

AUTHOR INFORMATION

Dr. Michael Marsh is a professor of Management Information Systems in the John L. Grove College of Business at Shippensburg University where he teaches a wide range of information systems, technology, and analysis courses at the graduate and undergraduate levels. His academic credentials include a B.S. in Mathematics, a M.S. in Computer Science, an MBA, and a doctorate in Management Science. Current academic interests include instructional technology, courseware development, and applied analytical modeling. Dr. Marsh has written and been the project director for several successful technology grants and serves on the University Emerging Technology Committee.

REFERENCES

1. Anderson-Cook, C. M. (1999). An In-Class Demonstration to Help Students Understand Confidence Intervals. *Journal of the American Statistical Association , 7* (3).
2. Cleveland, W. a. (1984). Graphical perception. *J of ASA* , 531-554.
3. Darius, P. L. (2007). Virtual Experiments and Their Use in Teaching Experimental Design. *International Statistical Review , 75* (3), 281-294.
4. delMas, R. C., Garfield, J., & Chance, B. (1999). A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning. *Journal of Statistics Education , 7* (3).
5. Gourgey, A. F. (2000). A Classroom Simulation Based on Political Polling to Help Students Understand Sampling Distributions. *Journal of Statistics Education , 8* (3).
6. Lock, R. H. (2001). *A Sampler of WWW Resources for Teaching Statistics*. Retrieved January 2008, from A Sampler of WWW Resources for Teaching Statistics: http://it.stlawu.edu/~rlock/maa51/onepage.html
7. Martin, M. A. (2003). "It's like... you know"; The Use of Analogies and Heuristics in Teaching Introductory Statistical Miethods. *Journal of Statistics Education , 11* (2).
8. Schwarz, C. J. (2003). An Online Sytem for Teaching the Design and Analysis of Experiments. *Joint Statistical Meeting,* (pp. 1-16). San Francisco.
9. Schwarz, C. J. (2007). Computer-Aided Statistical Instruction - Multi-Mediocre Techno-Trash? *International Statistical Review , 75* (3), 348-354.
10. Schwarz, C. J. (1997). StatVillage: An On-Line, WWW-Accessible, Hypothetical City Based on Real Data for Use in an Introductory Class in Survey Sampling. *Journal of Statistics Education , 5* (2).
11. Smith, G. (1998). Learn Statistics by Doing Statistics. *Journal of Statistics Education , 6* (3).

12. Vaughn, T. S. (2003). Teaching Statistical Concepts with Student-Specific Datasets. *Journal of Statistics Education , 11* (3).

13. Wechsler, H. L. (2000). Colleg Binge Drinking in the 1990s: A Continuing Problem. *Journal of American College Health , 48*, 199-210.

NOTES