Title

KICKING AND SCREAMING ABOUT STATISTICS: HOW SOCCER DATA CAN POTENTIALLY ALLEVIATE STATISTICAL ANXIETY

Authors

Sanchez, Juana
Minosa, Marela Kay Roque
Cook, Dianne
et al.

Publication Date

2013-03-26

KICKING AND SCREAMING ABOUT STATISTICS: HOW SOCCER DATA CAN POTENTIALLY ALLEVIATE STATISTICAL ANXIETY

JUANA SANCHEZ (Contact) jsanchez@stat.ucla.edu
MARELA KAY MINOSA
University of California, Los Angeles

JOHNNY MASEGELA
Statistics South Africa

DIANNE COOK
University of Iowa

Submitted for Publication

INTRODUCTION

Schools and teachers face the challenge of including more statistics in the K-12 curriculum (Cohen, 2012). However, the Common Core (CC) standards recommend less statistics at the elementary school level than is recommended by the GAISE report and NCTM standards. The pedagogy proposed by CC is also very different from Levels 1 and 2 of GAISE and the more general guidelines of NCTM (Bargagliotti and Webb, 2011). GAISE and many school curriculum reforms around the world aim to enhance the process of statistical inquiry across the curriculum from a very early age and to make young students statistically literate.

Fostering a positive attitude to statistics through using data that is both relevant and real to answer questions that are relevant to students using statistics does not need to involve only the teachers in the classroom. STEM education outside of formal school environments via Community/Youth after school family programming is getting increasing attention (Bell et al., 2009) in the United States. To improve understanding of the data gathering process, its purposes and benefits to society, taking advantage of students' interests outside the classroom to engage them in data gathering about themselves and bringing their data back to the classroom is an alternative that teachers could explore further. South Africa is doing that and, in this paper, we explore how their approach could be integrated into our school's path towards statistical literacy.

Inspired by South Africa as the host country for both the 2010 FIFA World Cup and the International Statistical Institute's (ISI) biennial meeting in 2009, Statistics South Africa (Stats SA) recruited the assistance of Johnny Masegela, otherwise known as Black Sunday during his prime as the Orlando Pirates' striker. Stats SA was convinced that the World Cup, as well as leaving a legacy in buildings and physical infrastructures, would present a meaningful, educational role for children of all levels, especially in mathematics

and statistics. Masegela also argued that statistics, mathematics and soccer could potentially have a mutually beneficial relationship. Together they created the innovative SOCCER4Stats program as the extracurricular activity for the Capacity Building Program of Stats SA, ISIbalo 6. The goal of SOCCER4Stats is to help individuals see "statistics from a different view other than just mathematics and to see how it can be applied to everyday events and activities". Today Masegela is traversing the country of South Africa, introducing these innovative activities to various schools as part of the ISIbalo 6. He continues to display how education, particularly statistics and mathematics, can gain from the soccer spectacular.

So we asked ourselves: could an innovative program like this help motivate school students in the United States to learn about statistics? Will working with real data like that collected by the SOCCER4Stats activities create an interest for both male and female students to become statistically literate? Can it be primarily about statistics, not only soccer? We set as our goal to assess the incidence of youth's participation in soccer in the U.S. and to assess venues through which SOCCER4Stats could be implemented as an extracurricular activity for active learning in the classroom. In this paper we explain what we found. We also explored further alternative ways in which the data collected through the SOCCER4Stats program can be used to illustrate important statistical concepts to students in the classroom.

Using data collected by students as a motivator to engage students in statistics learning is not new (Neumann et al., 2010). Carl Lee developed the "Real-Time-Online-Database," STATAC (Lee, 2013) whereby university students from different schools can answer several questions online and their data added to a large database and shared with other schools. This is an idea similar to that of CensusAtSchools (Davies et al. 2012), an internet-based project that uses awareness-raising about national censuses to involve schoolchildren in statistics. The databases collected from students are online and accessible. Those two projects encourage the use of real data, from and about school children, and promotes the teaching and learning of statistical thinking skills in the classroom. Based on the CensusAtSchool project, and inspired by the 2012 London Summer Olympic Games, SportsAtSchool (SportsAtSchool, 2013) was launched in 2011 to collect data from students via a questionnaire similar to that of the CensusAtSchool. Investigations of the data collected by those projects are presented in teacher resources based on the four-step statistical process as defined by the Guidelines for Assessment and Instruction in Statistics Education (GAISE) (Franklin et al. 2007) :

1. Formulate a question that can answered by data
2. Design and implement a plan to collect appropriate data
3. Analyze the collected data by graphical and numerical methods
4. Interpret the analysis in the context of the original question.

While the idea of CensusAtSchool and SportsAtSchools is similar to SOCCER4Stats for ISIbalo6, the collection of data in the former is done through a questionnaire. It works much like how the United States Census works – by filling out a survey. SOCCER4Stats, on the other hand, includes the students to partake in legitimate activities as part of their

data collection. The students are actively, in the literal sense of sports, partaking in the data.

Using sports data to motivate students to learn statistics is not new either. Albert (2003) popularized the idea of using Baseball data to teach statistics. Stephenson et al (2009) and Lock (1998) used a golf inspired game with the same purpose. Lock (2006, 2010) proposed teaching introductory statistics using sports examples. But these examples are based on tertiary data on player's performance compiled by others or on simulations, thus although students are integrating real-life data analysis in their learning of statistics those activities do not engage students actively in their own data collection process hence missing the opportunity to involve them in the whole statistical process.

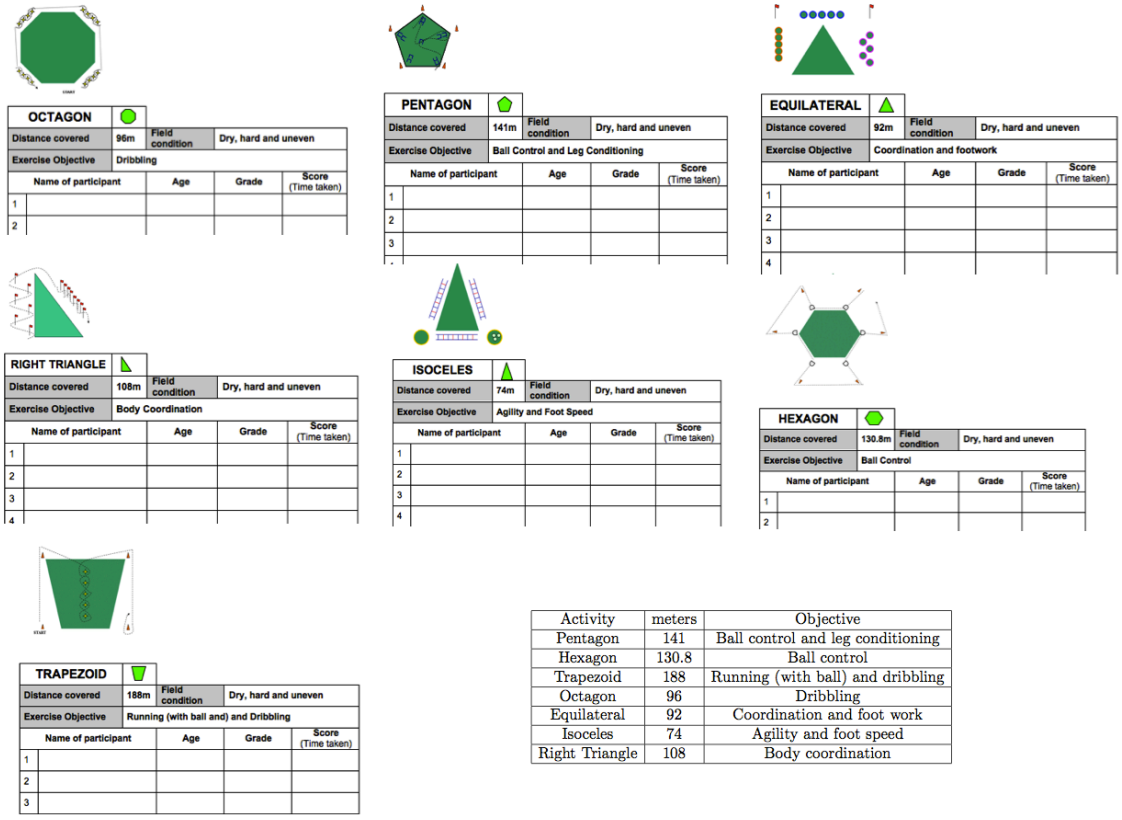
Although some statistics educators question the impact of active learning activities like those mentioned above on learning outcomes (Pfaff and Weinberg, 2009), many others (implicitly or explicitly) adhere to a constructivist approach to learning which requires a mode of learning that allows the learner to investigate freely, in realistic circumstances and meaningful contexts. If the students practice what they are learning by performing analyses of data they collect themselves, they gain deep understanding, which will then enable the student to properly apply what s/he has learned. This actually involves two requirements: personal investigation, and complex and meaningful context. (Libman, 2010).

In the rest of this paper, we describe the SOCCER4Stats activities, and exploratory analyses done in previous learning events which are appropriate for elementary school students. Following that, we describe what we found about the incidence of soccer interest in the United States and the potential geography within which using these activities might have most success. Finally, we present other classroom exercises with technology that are appropriate for middle, high school and even university students.

WHAT IS SOCCER4Stats?

At the 2009 International Statistical Institute's (ISI) biennial meeting held in South Africa, students from all over the world participated in the first International Statistical Literacy Project competition (ISLP, 2009) and a week of statistics activities tailored to students. As part of the extra curricular activities of the competition, Johnny Masegela took students to a soccer field near the ISI convention center and introduced some of the students to the SOCCER4Stats activities, with the help of South Africa's high school students that regularly played soccer (ISIBANE, 2009). Both the soccer players and the students participating in the ISLP competition did the activities and collected data. After the activities, educators from the United States and South Africa, in a classroom equipped with a computer for each student, exposed these students to modern statistical graphing techniques and new statistical methods using the data collected in the activities. With what they had learned, the students then presented their analysis to the large audience of statisticians attending the ISI meeting.

The activities of SOCCER4Stats can be briefly described as follows: In a soccer field, students draw the circuits indicated in Figure 1 and complete each circuit. Each circuit has a purpose. For example, the isosceles sketch helps students gain agility foot speed. The different exercises test the players' coordination, control, conditioning, dribbling and speed. "This is important in the development of soccer skills," says Jonny Masegela. While a student runs the circuit with the ball, another student times the speed at which the player completes the activity and other variables, writing results in spreadsheets like those displayed in Figure 1. At the end of the activity, there is a spreadsheet for each circuit completed for all students participating. The data is then put together in a combined Excel spreadsheet. The students then take the data to the classroom and analyze it.



Activity	meters	Objective
Pentagon	141	Ball control and leg conditioning
Hexagon	130.8	Ball control
Trapezoid	188	Running (with ball) and dribbling
Octagon	96	Dribbling
Equilateral	92	Coordination and foot work
Isoceles	74	Agility and foot speed
Right Triangle	108	Body coordination

Figure 1. SOCCER4Stats data collection activities by students.

At the ISI 2009, Dianne Cook discussed with students several potential research questions and discussed how the questions could be answered by use of statistical graphs and the data that the students collected. For example, with graphs shown in Figure 2, Dianne Cook showed students how to compare the speeds in all activities of two groups of students in a snapshot. The graphs showed that the students who were soccer players were more consistent and faster. Then because students were interested in finding out who the fastest was, similar plots were drawn by player (not included here). This brought the discussion as to what was more appropriate to use: the mean, or the median? It turned

out that the ranking of the players by speed varied considerably depending on whether the mean or the median was used.

In addition to the above analyses, Dianne Cook showed students how they could look at the relation between age and speed and how that relation changed across several well defined subgroups. This entails looking at two factors: group (with two levels, statisticians or soccer players) and gender (also with two levels, male, female). Figure 3 illustrates how she used ggplot to do that and the conclusions extracted from the data.

The SOCCER4Stats activities were good motivators for students to understand the difference. Other questions that came up were: Who is the most consistent? Which activities took longer? Were some activities easy for everyone, or were there any that had a lot of variation from athlete to athlete? Are older kids faster? Does grade make a difference? How do the two girls fare compared to the boys?

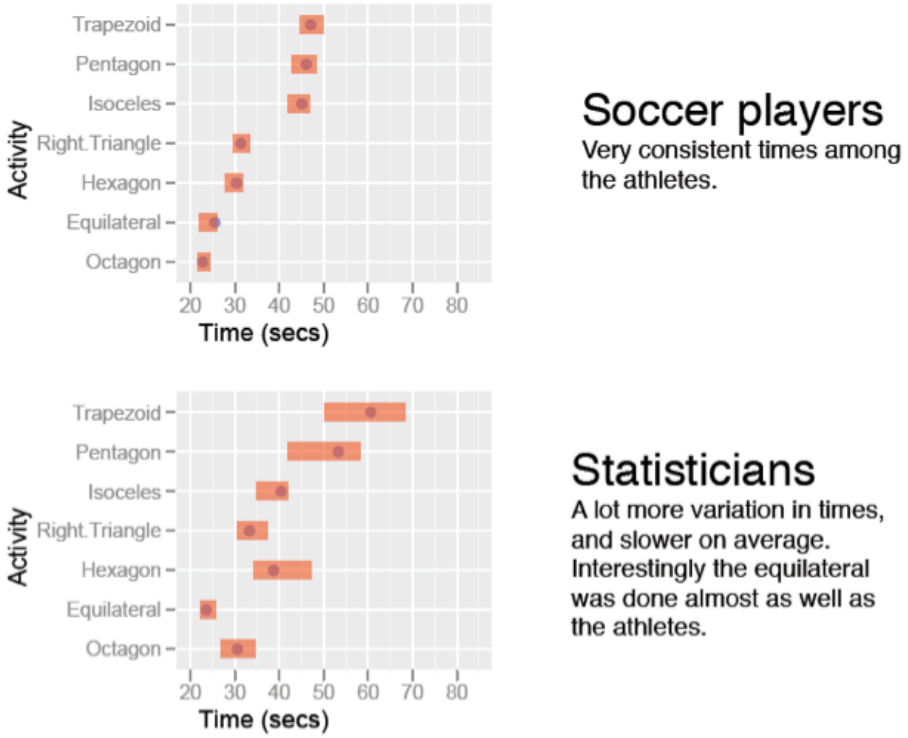


Figure 2. Comparing two groups of students using ggplot.

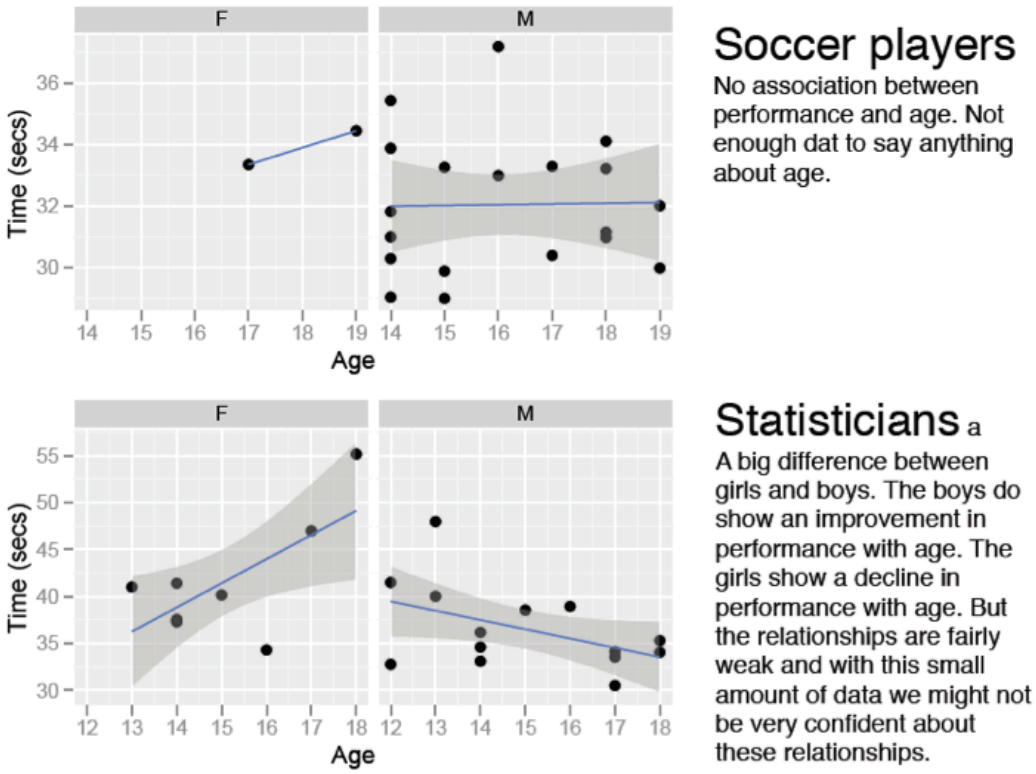


Figure 3. Relation between age and speed for subgroups of students.

After a tutorial by Dianne Cook and Naomi Robbins, students used their computers and Excel to do their own analyses of the data. Figure 4 illustrates one of their analyses.

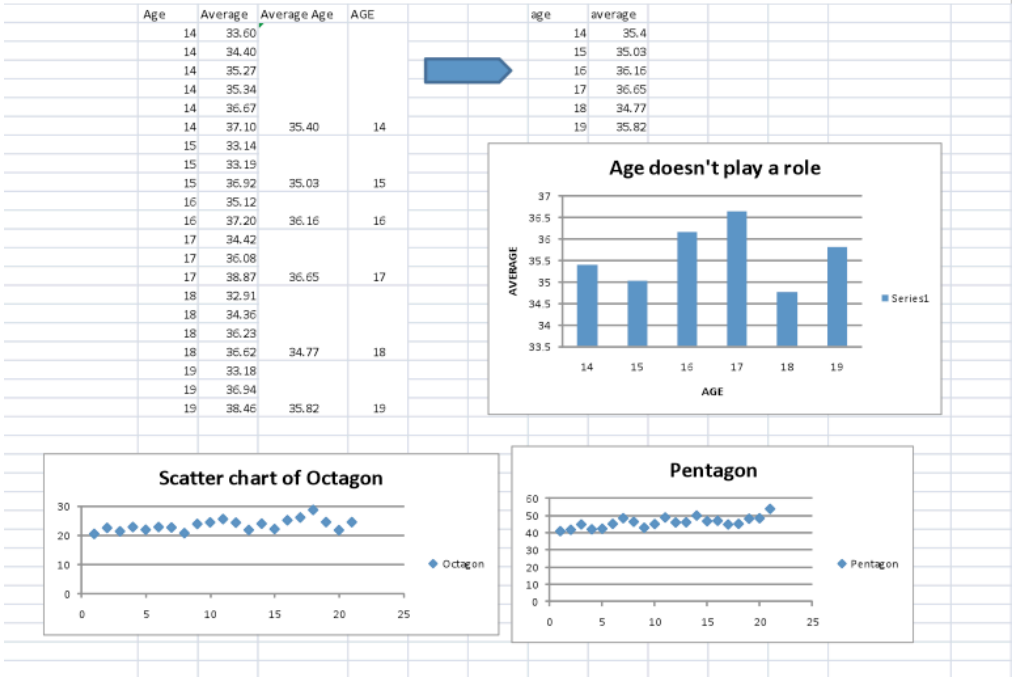


Figure 4. Example of analysis done by students with Excel.

The SOCCER4Stats activities not only engage students in data collection on their own performance. They also challenge them to think about potential variables that might help explain soccer outcomes to help soccer scientists answer important research questions. Thus, the data helps students be involved in scientific enquiries conveying to them the notion that statistics is done to answer questions across the curriculum.

WHY SOCCER?

Statistics of soccer are not as well known as baseball or other sports data. Kaplan (2013) explains how people in soccer have historically paid little attention to statistics because the sport does not lend itself to statistical scrutiny. It is difficult to quantify a player's performance, as it involves the interaction with all the other players' passes and shots. Blalik (2008) claims that the very nature of soccer seems to defy statistical analysis and soccer scientists haven't yet identified new statistics that correlate best with winning matches. "Part of the problem is that hundreds of events — and 20 players running about 10 kilometers each — lead to just a handful of goals, if any, in each match. In many cases, it's hard to ascribe credit or blame for those goals to individual actions." Thus, soccer presents itself as a scientific question that students need to solve not only with statistics but with knowledge they have from other areas of the curriculum, and their own experience in the field.

Not all students in our schools like soccer, thus what could be the incidence of implementing SOCCER4Schools in schools in the United States? How widespread is the interest in soccer in the United States?

According to the U.S. Census Bureau, in 2009 approximately 13.5 million people in the United States participated in soccer. This number is based on a questionnaire distributed to 10,000 households which represented approximately a population of 270 million individuals 7 years of age or older who participated in various sports more than once that year. The reported 13.5 million persons is about 5% of that population. Of that 5%, approximately 60%, or 8 million people, were considered youth (ages 7 to 17.)

Now why might participation in soccer be largely attributed to the youth, along with other sports? In order to combat obesity in today's youth, First Lady Michelle Obama created the program "Let's Move" to help the youth of today follow a "path to a healthy future" (Let's Move). Similar to her program, there have spawned various youth sport organizations focused on promoting a healthier and more active lifestyle for today's youth. Perhaps even the sport of soccer is no exception. With the creation of youth soccer programs such as the United States Youth Soccer Association (USYSA), the American Youth Soccer Organization (AYSO), and more, it is not surprising that more of today's youth are motivated to participate in soccer. So whilst the sport of soccer was first documented at the collegiate level around 1840, the development of interest garnered for the sport begins with the youth (Handley et al, 1994).

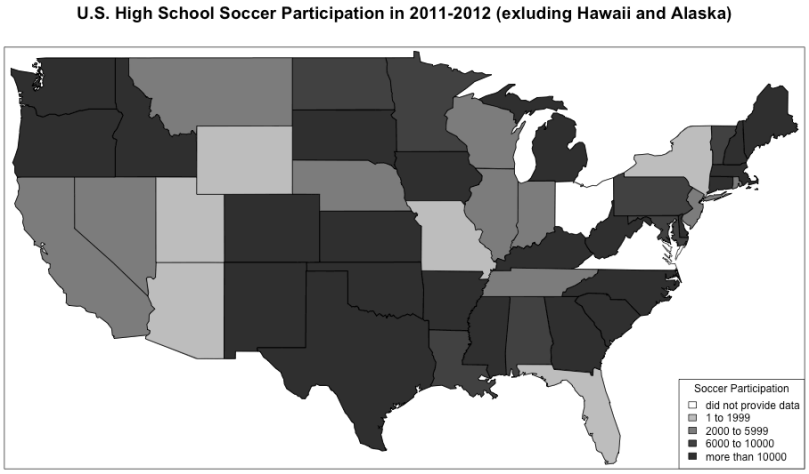


Figure 5. U.S. High School Soccer Participation.

As conveyed in the map of U.S. High School Soccer Participation of Figure 5, soccer participation at the high school level is spread throughout the United States. This information was taken from the National Federation of State High School Associations' (NFSHSA) Participation Survey 2011-2012. The survey was mailed to all 51 of their associations in the United States with sports teams competing at the high school level (all 50 states with the addition of the District of Columbia). It is important to note that the data from this survey is voluntary, so it is understandable to have high schools not respond, which would explain the reported 0 participants for two states. So while the concept of the survey is ideal, it lacks support. Nevertheless the data provided conveys a general interest in competitive soccer for high school students throughout the United States.

In addition to soccer being a sport played amongst the youth, it has something most major sports in the United States lack – a relatively high number of female participants. According to the U.S. Census Bureau, soccer is amongst the top ten most popular sports for both males and females based on the number of participants between 2009 and 2010 (U.S. Census Bureau, "Table 1248" 2012). In 2010, there was a reported estimate of about 12% female participation (23,650 of 186,460 female participants) in the NCAA Sport soccer as compared to the approximately 8% male participation (21,770 of 249,307 male participants) (U.S. Census Bureau, "Table 1247" 2012). Whilst soccer is still thought of as a predominantly male sport, there is a nearly equivalent amount of participation from females with a reported proportion of 53% males and 47% females by the NFSHSA.

Interest in soccer is widespread in youth and growing, as we have seen. Thus, if not of interest to all students, learning by the use of soccer activities and data collected from students with SOCCER4Stats can potentially create an interest in statistical language for both girls and boys. It could also potentially bring the extracurricular activities of students and those at school engaged in them (such as, for example, the Soccer coach) into the classroom, and create interest in other parts of the curriculum.

POTENTIAL ADDITIONAL ACTIVITIES FOR THE CLASSROOM

Because the data collected in South Africa with the SOCCER4Stats activities is observational, and was not planned to be collected in an experimental setting, the analysis done does not allow to extract causal conclusions. However, the data can be used to conduct more advanced statistical analysis such as regression, which would be appropriate for high school and middle school students.

Teachers may divide the class into two groups – half of the students will be considered data collectors, while the other half will be the players for the data by participating in the seven SOCCER4Stats activities. This can be done by one of two ways: fixing the students to two separate roles or by having all students experience both roles - data collector and player. Then with each activity, they will be able to measure different qualities of the soccer players. With the measurements they will then input the names, ages, grade levels and scores (time taken to complete activities) into the spreadsheets provided by SOCCER4Stats. Once the collection is complete, as a class, the students can employ basic statistical concepts, such as correlation, significance, and linear regression, on what they gather. They will be able to discuss as a group what kind of statistical questions they could ask from their data and answer them in meaningful ways.

For example, for high school students, a topic within the course could be teaching the basics of linear regression. Essentially the students would be introduced to how they could potentially predict a single outcome by explaining its relationship to explanatory variables. An example of this was done on data collected from the SOCCER4Stats program in ISI 2009 and additional data sent in by Johnny Masegela (the data is available upon request). The following Table 1 explains the variables used from the data for linear regression. The data was collected on 90 students participating in the program. It consisted of continuous variables – the ages, the grades, time taken for each activity (in seconds), and the number of errors they make – and discrete variables – when they participated (grouped by period), what role they played, the genders, and the activities they participated in (grouped by shape). With this data we are interested in the time taken by the participants. Because the activities vary, we standardized the time taken by dividing by the standard deviation – 15.549.

Variable Name	Description
role	discrete (newplayers, player, statistician, and test data): the role of the students.
age	continuous: the ages of the participants (ranging from 12 to 19 years).
grade	continuous: the grades of the participants (ranging from 6^{th} to 12^{th}).
gender	discrete (male and female): the gender of the participant
activties	discrete (the 7 SOCCER4Stats activities): the activity from which the time was collected.
time	continuous: time relative to each activity for each participant; outcome variable.
errors	continuous: the number of errors made in particular activity

With this particular example, we utilize a backwards, stepwise regression – begin with a full model and remove variables to reduced model. Now when performing regression, there are four assumptions that must be fulfilled for the results to be significant: 1) normality, 2) independence, 3) homoscedasticity, and 4) linearity. In regards to normality, the variable of interest (our outcome variable, time) must be normally distributed. When plotted, its density should resemble a bell-shaped curve. To fix this possible problem, we transform the standardized time data by a log transformation.

$$\log(\widehat{st.time_i}) = \beta_0 + \beta_1 grade_i + \gamma_1 statistician_i + \gamma_2 testdata_i + \varphi_1 male_i + \alpha_1 pentagon_i + \alpha_2 equilateral_i + \alpha_3 right.triangle_i + \alpha_4 isoceles_i + \alpha_5 hexagon_i + \alpha_6 trapezoid_i + \varepsilon_i$$

	Df	Sum Squares	Mean Squares	F-Value	Pr(>F)
role	3	11.236	3.7455	79.4917	<0.001
age	1	0.050	0.0497	1.0542	0.305
grade	1	0.018	0.0177	0.3763	0.5398
gender	1	1.850	1.8502	39.2664	<0.001
activity	6	34.441	5.7401	121.8248	<0.001
Residuals	617	29.072	0.0471		

Now we can continue with the regression; we will begin with a full model. Next with partial f-tests, we'll select the variables. Given the full model, we performed an F-test to check for significant differences in means. As conveyed in the previous table, the explanatory variables age and grade are not significant as denoted by their p-values being greater than 0.05. Because they do not seemingly provide meaningful information for our model (any affect could be due to chance), we remove them to create the reduced and final model. Finally with all things considered, our final model uses the role of student, their gender, and the activity they participated in as predictor variables for the log transformed standardized times.

$$\log(st.\widetilde{time}_i) = 1.111 - 0.213(player_i) - 0.119(statistician_i) + 0.232(testdata_i) - 0.139(male_i) + 0.229(pentagon_i) - 0.434(equilateral_i) + 0.092(right.triangle_i) + 0.011(isoceles_i) + 0.030(hexagon_i) + 0.364(trapezoid_i) + \varepsilon_i$$

In the end with the final model, we can interpret that with all things constant, a female new player participating in the octagon activity will complete the activity in approximately 47.23 seconds. Comparatively, if the student were to be a statistician instead of a player, the approximated time to complete the activity would decrease by 5.30 seconds. To understand what we want to know from the transformed, standardized times, we simply take the intercept least squares estimate, with the addition or subtraction of the factor variables and multiply that by the standard deviation of the initial times. For instance, with all things constant, a male, statistician participating in the trapezoid activity would complete the activity in approximately 52.51seconds.

CONCLUSIONS AND FURTHER RECOMMENDATIONS

Other potential analyses that could be done with the SOCCER4Stats activities may be dictated by the topic at hand being discussed in the classroom and the interest of the students. Similarly, other variables could be collected, and an experiment could be planned, such as allocating students randomly to do the activities with a prize at the end for the fastest, and other students being allocated to the activities but without a prize.

Teachers could also compare the test scores at the end of the quarter of those students who took part in the SOCCER4Stats program, which obviously should be volunteer, and those who did not.

The main point is that all the areas covered by the Common Core standards can be studied with the SOCCER4Stat activities while keeping up with the GAISE and NCTM standards and the ideal enhancing the process of statistical inquiry across the curriculum from a very early age and to make young students statistically literate.

REFERENCES

Addona, Vittorio. (2010). Using Sports Data to Motivate Statistical Concepts: Experiences from a Freshman Course. *International Association of Statistical Education*.

Bargagliotti, A.E. and Webb, D. (2011) Elementary School Teachers: Teaching, Understanding, and Using Statistics. *Statistics Teachers Network*, Winter 2011, p. 5-10.

Bell, P., Lewenstein, B., Shouse, A.W. and Feder, M. A., Editors (2009). Learning Science in Informal Environments: People, Places, and Pursuits.Committee on Learning Science in Informal Environments, *National Research Council. The National Academy Press.* (http://www.nap.edu/openbook.php?record_id=12190).

Blalik, C. (2008). Can Statistics Explain Soccer? New York Times, June 24.

Cohen, J. (2012). The Common Core Standards. Where do Probability and Statistics Fit in? *Statistics Teacher Network*, Spring 2012. P.2-5. http://www.amstat.org/education/stn/pdfs/stn79.pdf

Davies, N., Richards, K., Aliaga, M., and Nichols, R. CensusAtSchool. *Significance.* December, 2012, p 175. http://www.amstat.org/censusatschool/

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. 2007. *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework*. Alexandria, VA: American Statistical Association.

Wickham, H. *GGplot2. Elegant Graphics for Data Analysis.* Spriner-Verlag, 2009.

Handley, A., Handley, C., Handley, H., Handley, L. M., Handley, L. R., Handley, N. (1994). "Youth Soccer in the United States". *The Geographical Bulletin.* 36(1).

ISIBANE (2009). ISIBANE Newsletter 2, page 1. http://www.statssa.gov.za/isi2009/ISIbane_Newsletter_2.pdf

ISLP (2009) http://www.stat.auckland.ac.nz/~iase/islp/competition-first_winners

Kaplan, T. (2010). When It Comes to Stats, Soccer Seldom Counts. New York Times, July 8th.

Lee, C. STATACT. http://stat.cst.cmich.edu/statact/

Let's Move. http://www.letsmove.gov/learn-facts/epidemic-childhood-obesity.

Libman, Z. (2010). Integrating Real-Life Data Analysis in Teaching Descriptive Statistics: A Constructivist Approach. *Journal of Statistics Education* Volume 18, Number 1 (2010),

Lock, Robin H. (1998). Using Simulation, Sports, and the WWW to Help Students Experience Experimental Design. *Statistics Education Research Journal.*

Lock, Robin H. (2006). Teaching an Introductory Statistics Class Based on Sports Examples. *Statistics Education Research Journal.* 7.

Lock, Robin H. (2010). Statistical Models for Student Projects With Sports Themes.

Statistics Education Research Journal.

Neumann, D. L., Neumann, M. M., and Hood, M. (2010). The development and evaluation of a survey that makes use of student data to teach statistics. *Journal of Statistics Education*. 18(1).

Pfaff, Thomas J. and Weinberg, Aaron. (2009) Do Hands-On Activities Increase Student Understanding?: A Case Study. *Journal of Statistics Education*. 17(3).

SportsAtSchool (2103) http://www.censusatschool.org.uk/take-part/questionnaires/sportatschool-20112012

Stephenson, P., Richardson, M., Gabrosek, J. and Reischman, D. (2009). How LO can you GO? Using the Dice-Based Golf Game GOLO to Illustrate Inferences on Proportions and Discrete Probability Distributions. *Journal of Statistics Education*. 17(2).

The National Federation of State High School Associations. "2011-12 High School Athletics Participation Survey". *(NFSHSA).*

U.S. Census Bureau. (2012). "Table 1247. Participation in NCAA Sports by Sex: 2009 to 2010."

U.S. Census Bureau. (2012). "Table 1248. Participation in High School Athletic Programs by Sex: 1980 to 2010."

U.S. Census Bureau. (2012). "Table 1249. Participation in Selected Sports Activities: 2009".