USING PERSONAL ACTIVITY DATA IN AN UNDERGRADUATE STATISTICS COURSE

Kady Schneiter[1], Lacy Christensen[1], and Victor R. Lee[2]
[1]Department of Mathematics and Statistics
[2]Department of Instructional Technology and Learning Sciences
Utah State University, Logan, Utah 84322
kschneiter@usu.edu

Personal activity data (PAD) obtained from activity trackers has the potential to stimulate thinking about statistics in a way that other forms of data, even other real data, cannot. Because the data come from the students' own activities, they are intimately familiar with them and able to reason about patterns and variations in the data based on their own experience. We describe our experiences using personal activity trackers (specifically Fitbit activity tracking devices) to engage students in statistical thinking in an undergraduate statistics course and discuss our findings about how PAD can be naturally integrated into the introductory statistics curriculum to promote student interest and engagement, and to facilitate statistical thinking.

INTRODUCTION

Many authors have discussed the importance of using real data in statistics instruction. Real data are interesting and engage students in the learning process (Libman, 2010; Neumann, Hood and Neumann, 2013); real data are often messy and can provide students with opportunities to develop data-management skills (Carver et al., 2016; Carver and Stephens, 2014). Cobb and Moore said "To teach statistics well, it is not enough to understand the mathematical theory; it is not even enough to understand also the additional, non-mathematical theory of statistics. One must, like a teacher of literature, have a ready supply of real illustrations, and know how to use them to involve students in the development of their critical judgment." (1997, pg. 803)

Sources of interesting datasets abound on the internet, however, statistics educators can go one step farther in making data relevant and interesting to their students by enabling them to collect it themselves. Modern social trends, such as the quantified-self movement (see Lee, 2013) have generated excitement about using personal activity data while trends in technology, specifically wearable devices, facilitate the collection of large amounts of student-generated data. The business of wearable technology is booming. 104.3 million wearable devices were shipped in 2016 and industry analysts forecast that the number will reach 125.5 million this year and 240.1 million in 2021 (Lamkin, 2017). This proliferation of data provides opportunities for students of statistics. Devices that record activity data facilitate the collection of personal data automatically. Such data are not only real but are relevant. Students are intimately aware of the background and context of such data. Statistics educators have the opportunity to explore how to take advantage of such data in statistics courses. Lee, Drake, and Thayne (2016) have described using activity tracking devices to foster statistical thinking in elementary students but less has been done at the undergraduate level. Thayne (2016) describes his work using data from activity trackers with students enrolled in an undergraduate statistics course. His work was done with individual students rather than with regular course activities. In this paper, we describe our experience using Personal Activity Data (PAD) obtained via activity tracking devices to engage undergraduate students in the processes of learning about and doing statistics. The main questions of interest to us are 1. Where can PAD activities be most effectively used in an introductory statistics course? 2. How do students respond to the use of PAD in statistics activities?

BACKGROUND

We conducted our study during the summer of 2017 in an undergraduate, calculus-based statistics course for scientists and engineers at a public university. Twenty-six students enrolled in the course; of these, 19 were male and the rest female. Students came from a variety of majors including Computer Science, Wildlife Resources, Geology, and Mathematics. Most students took the class as a requirement for their majors.

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

We chose to run the study during the summer in order to be able to work closely with a small group of students. During the regular school year, the course runs with 150 or more students per semester. We plan to use the personal data in the large sections of the course in the future.

We obtained IRB approval before the start of the study. A student researcher described the objectives, procedures, and possible risks of the study to the students before inviting them to participate. Students were given the option of working with data generated by the student researcher rather than generating their own data, however, all 26 students enrolled in the course chose to participate in the study and collect their own data. All students completed at least one of the tasks associated with the study; 19 completed all relevant work.

Each student was issued a Fitbit Flex (see Figure 1) wristband for the duration of the course. These devices track numbers of steps taken, distance traveled, calories burned, hourly activity, and sleep (see Figure 2). We chose Fitbits over other possible devices because of their availability and the accessibility of the tracked data. Fitbit data can be easily recorded by syncing the device with an app available for smartphones, tablets, or computers. When using a smart phone, the Fitbit will sync automatically and wirelessly. The user can choose to do this by explicitly telling the Fitbit to sync or can set the device to sync continually.

Students were issued the activity trackers on the first day of class and were taught to sync their activity information with a device and access their data. The devices were each assigned a number and data were then associated with these numbers rather than with student names to protect privacy. Students were requested to wear the device at all times, except when bathing or charging.

Figure 1: A Fitbit Flex

Figure 2: The smartphone app for the Fitbit.

The class met for 70 minutes per day, four days per week over a seven week period. Three of the four days each week were spent in lecture and other regular class activities. On the fourth day, we engaged in application tasks during which students collected and worked with real data. Four of these tasks used the personal activity data collected from students. For each task involving the activity data, the student researcher accessed the data for each Fitbit, using a custom data acquisition tool described in Lee et al (2016) and made it available to the students through our online learning management system, Canvas. Students examined their own data for each of the relevant tasks.

MOTIVATION

Using data from activity tracking devices provides opportunities in line with the recommendations of The Guidelines for Assessment and Instruction in Statistics Education

(GAISE) report endorsed by the American Statistical Association. These recommendations include directives to “Integrate real data with a context and purpose,” to foster active learning, and to use technology to explore concepts and analyze data.

In their paper titled, Improving the Teaching of Applied Statistics: Putting the Data Back into Data Analysis, Judith singer and John Willett (1990) discuss desirable characteristics of real data sets used in statistics teaching. These characteristics are authenticity, background information, interest and relevance, substantive learning, availability of multiple analyses, the importance of raw data, and case identifiers. Using activity trackers to collect activity data provides students with data that possesses all of the characteristics. The data are authentic in that they are “actual measurements taken on an actual sample of cases”. The background information is not only available but is supplied by the students themselves. They know the story behind the data; they are able to recognize whether a peak in steps taken represents a long hike or a busy day at work. For these same reasons, the activity data is interesting and relevant to students. The data provide substantive learning and highlight the usefulness of statistical analyses helping students to observe and predict patterns of behavior in their own lives. The data present opportunities for multiple analyses. Though the tasks all involved the same variables, we approached them from a variety of angles: using descriptive statistics, understanding distributions, bivariate analysis, and inference. It is likely that there are other avenues of analysis that we have not yet tried. The data were given ‘raw’ to the students, just as they were taken from the devices so student had to deal with outliers and inconsistencies. Finally, case identifiers were naturally included with the data and related to the students’ own knowledge of what had occurred.

PAD TASKS

In our efforts to address the question ‘Where can PAD activities be most effectively used in an introductory college course?’ we developed a fairly diverse array of tasks. The data clearly lent themselves to investigation with descriptive statistics and regression analysis. We hoped to leverage familiarity with the data to facilitate thinking about distribution in a task on random variables. We used the context (but not, at first, the data) in an activity in which students designed an experiment to address a PAD related question. Later in the semester, for additional credit, some students carried out the experiment they had designed for the earlier task.

The first PAD task was conducted in the first week of school. At this point, little data had been collected so the task required students to use only the context of the data to design an experiment to compare mileage as tracked by the activity tracking device and mileage as recorded by a treadmill. Students were not required to gather data or provide an answer to the question at that time – they were only asked to describe how they would do so. Later in the semester, once methods of analysis had been discussed, students were given the opportunity to revisit and answer the question by conducting the experiment they had described.

The second PAD task required that students used data summary techniques to make sense of their data. They were asked to identify the waking times during which they were least active based on the consideration of at least two of the variables tracked by the device. Students were asked first to predict the answer to the question then investigate their personal data using graphical and numerical summary techniques. They were instructed to explain their choice of methods and what they learned from each summary. Finally, they were to state and justify their conclusions.

For the third PAD task, we used personal activity data to help students to understand the concept of random variables. The ideas of random variables and distributions are intrinsically connected. We chose to use random variables that described the number of steps a student had taken in a randomly chosen hour. Our idea was that students’ intimate knowledge of their activity patterns would enable them to describe the distribution of the random variable and thus to connect with the idea of distribution in a broader sense. After considering the distribution of the number of steps a student had taken in one randomly chosen hour, we asked them to consider how that would compare to the distribution of the number of steps taken in a second randomly chosen hour. In this way, we hoped to point students toward the idea of identically distributed random variables or the concept of sampling two values from the same population. Students used software to produce histograms of the distributions explicitly and compared their predictions to the actual distributions.

The activity data are a natural fit for an investigation of bivariate data. For the fourth PAD task, students investigated the relationships between variables tracked by the device in order to find the one that would be the best predictor of ‘calories burned’. Students first examined the strength of correlations between various variables and ‘calories burned’. Once they had chosen a predictor variable, they investigated methods of identifying a ‘line of best fit’ considering at least one method other than the least squares regression line. They made comparisons between these methods and discussed the interpretation in the context of the work they had done.

Finally, students were given the opportunity to carry out the experiment they had proposed in the first task to determine whether there was a difference between distance travelled as tracked by the device and that tracked by a treadmill. This task was optional but required students to refine their experimental design, gather data, choose appropriate methods for data analysis, and write up their conclusions.

STUDENT RESPONSE

Following each task, students completed a feedback form addressing their experience. One of the prompts, “Describe how the PAD affected your learning and interest in the task”, elicited a number of responses related to the general themes that emerged from the surveys. These themes are 1. Examining personal data added interest to the activities, 2. Using personal data had an influence on students outside of the class activities, and 3. Students perceived that using PAD affected their understanding or their efforts to understand the statistical principles we introduced in class. In the paragraphs that follow, we discuss each of these themes and illustrate them with typical student responses to the stated prompt.

When students examine personal activity data, they are attuned to patterns and trends in the results and can make sense of anomalies in a way that may not be possible with other types of data. As one student said of application task two, “It gave real world data that pertained to my life to analyze making it a lot easier to relate to the data”. A student who sees a large jump in the number of steps they took on a particular day can reflect on their activities to determine the source of the outlier. Perhaps, the student went for a long walk on the day in question. On the other hand, perhaps they will recognize that they did not take a large number of steps but they did mow the lawn and could possibly conclude that the vibration of the lawnmower interferes with accurate step counting. Thus the data have meaning and relevance in a way that no pre-packaged dataset, even a real dataset could have. As another student said of the fourth task, "I liked using PAD because it made me more interested in the outcome of the data. It was also easy to see if something wasn't right because I knew what the data should look like.”

Students’ comments also showed that they connected their work with the data to their personal lives. Of the second task, a student said, “I was more interested in the task because it was my own data. I am able to see how active I am and possibly make changes if necessary.” Another student commented about task four that “It was a lot cooler to see my own data because I can really see how much I move and I can improve myself.” Connecting the work students do in a statistics class to their daily lives can help them to see the relevance and usefulness of the discipline.

Although, for this study, we did not explicitly address how utilizing the personal activity data affected student learning about statistics, focusing instead on how to use PAD and how students respond to it, we found that students *perceived* that using these data did improve their learning. Said one student “The PAD added a personal dimension to the task which in turn kept me more interesting in the assignment and thus allowing me to learn more along the way.” Students themselves felt that by being more engaged in the process of statistical investigation they were able to understand more and even that they were motivated to try to understand more. As another student commented, “Using my own data allowed me to connect to the task and be more interested in doing the analysis correctly.” The students were invested in the analyses because the results had personal meaning and were consequently motivated make sure they were able to do them correctly. They found that connecting the statistical topics to practical and personally relevant questions enabled them to understand better how the methods should be applied.

While the students’ comments were largely positive, the activity trackers were not universally popular. One student in particular told the instructor that he felt like the device was ‘judging’ him.

Though no particular activity goal was imposed or even implied, this student created one for himself and felt guilty when he didn't get enough activity in a particular day. He was very glad to return the device at the end of the term. A similar concern, in which the enjoyment of a physical activity was diminished by the fact that it was being tracked was reported by Lee and Drake (2013). We think it is important to avoid imposing a specific activity goal on students. Such an approach may cause stress and is beyond the purpose of the course. Rather, we emphasized the goal of exploring patterns of activity or inactivity, whatever they are. Other issues related to the technology also frustrated students. For instance, a few had repeated problems with syncing their devices and so never had enough of their own data to work with; others found it uncomfortable to have to wear the wristband.

CHALLENGES

When we were organizing the course content, we decided we wanted to spread the use of the devices over the semester. We thought that this would keep students interested and give them perspective on trends and patterns in the data. We conducted the data summary task, the first to use the device data, during the second week of classes. Though students were excited to use their devices, the task would have been more effective had we left it to later in the semester. At the second week, little data had been collected and some students had been unable to sync their devices and so had no data at all. These students were instructed to use data collected by the student researcher, but without their own data, the task lost much of its interest. Tasks involving the PAD would be better left until problems with data collection had been worked out and more data were collected.

The data summary task was also complicated by the analysis software that we used. Students use R for analysis in this course. Students were still very new to R when we engaged in this task and so their ability to complete the task well was complicated by the novelty of the software. This issue could also have been resolved by assigning the task later in the semester or investing more time early on in helping students to become proficient with the software.

CONCLUSION

Wristbands and other wearable technologies present statistics educators with an opportunity to engage students in collecting and analyzing personal data. Wearable devices facilitate the collection of real data with little effort on the part of the student; data that is relevant to students and with which they are inherently familiar. We were able to design lessons to take advantage of the PAD that were relevant to a variety of important statistical topics. Our tasks related to experimental design, data summary, random variables, bivariate data, and inference and there are likely other topics with which the data could be effectively used. We found that the majority of students were excited to work with data that they had generated themselves. Many reported that using the PAD helped them to be engaged in the tasks and some commented that they felt motivated by the personal nature of the data to understand better the statistical topics.

REFERENCES

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., ... & Wood, B. (2016). Guidelines for assessment and instruction in statistics education (GAISE) college report 2016. *Alexandria, VA: American Statistical Association.* Retrieved from www.amstat.org/education/gaise.

Carver, R. H., & Stephens, M. (2014). It is time to include data management in introductory statistics. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iaseweb.org/icots/9/proceedings/pdfs/ICOTS9_C134_CARVER.pdf .

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American mathematical monthly*, *104*(9), 801-823.

Lamkin, P. (2017) Wearable Tech Market to Double by 2021. Retrieved from https://www.forbes.com/sites/paullamkin/2017/06/22/wearable-tech-market-to-double-by-2021/#3e454791d8f3.

Lee, V. R. (2013). The Quantified Self (QS) movement and some emerging opportunities for the educational technology field. *Educational Technology*, (November-December 2013), 39.

Lee, V. R., & Drake, J. (2013). Digital physical activity data collection and use by endurance runners and distance cyclists. *Technology, Knowledge and Learning, 18*(1-2), 39-63. doi:10.1007/s10758-013-9203-3

Lee, V. R., Drake, J. R., & Thayne, J. L. (2016). Appropriating Quantified Self Technologies to Improve Elementary Statistical Teaching and Learning. *IEEE Transactions on Learning Technologies, 9*(4), 354-365. doi:10.1109/TLT.2016.2597142

Libman, Z. (2010). Integrating real-life data analysis in teaching descriptive statistics: A constructivist approach. *Journal of Statistics Education*, *18*(1), 1-23.

Singer, J. D., & Willett, J. B. (1990). Improving the teaching of applied statistics: Putting the data back into data analysis. *The American Statistician*, *44*(3), 223-230.

Neumann, D. L., Hood, M., & Neumann, M. M. (2013). Using real-life data when teaching statistics: student perceptions of this strategy in an introductory statistics course. *Statistics Education Research Journal*, *12*(2).

Thayne, J. (2016). *Making Statistics Matter: Using Self-data to Improve Statistics Learning* (Unpublished doctoral dissertation). Utah State University. Logan, UT.