Research Statement

Karsten T. Maurer

My time as a graduate student at Iowa State University has afforded me many opportunities for research. Assistantship work, coursework and numerous small projects have presented me with the experience of a wide sampling of research topics from which I have cultivated a core set of academic interests. My current body of research is in undergraduate statistics education and statistical graphics. Both fields offer exciting, interesting and important research opportunities. Undergraduate statistics education is rapidly expanding, in large part through service courses offered for other majors and disciplines. This gives us the chance to spread statistical literacy to a much larger population than ever before and research into improving statistics education ensures that we make the best use of this opportunity. Statistical graphics play an integral role in data exploration, data display and model diagnostics; a role that only becomes more interesting and challenging as data sources grow in complexity and size.

Within undergraduate statistics education I am focused on researching how technology can be employed to improve student comprehension of statistical concepts. An exciting area that is rapidly developing is the use of computationally intensive randomization-based methods for teaching statistical inference. This approach teaches statistical inference through tools such as bootstrap confidence intervals and permutation/randomization hypothesis testing. The idea is that we can introduce students to the concepts of statistical inference without the requisite cost of first learning to work with theoretical probability distributions. A major portion of my thesis research has been in leading a study that compared learning outcomes between theoretically based and randomization-based curricula for statistical inference in a designed experiment. The study randomly assigned students of an introductory statistics course to one of two rooms, each receiving one of the teaching approaches during the unit on statistical inference. The classrooms were taught using a co-teaching structure to avoid confounding the curricula effect with the instructor effect. Using a model based approach we found that learning outcomes for confidence interval topics were improved significantly for students receiving the randomization-based inference curriculum. In future research, I would like to develop methods for curricula assessment and also further explore randomization-based teaching methods. A specific idea would be to investigate if using tactile randomization (e.g. shuffling cards, flipping coins, etc.) in the randomization-based inference curriculum would improve learning outcomes as compared to the use of computer randomization alone.

Another area of my research has been in the development of educational tools that connect undergraduate students to large data sources. Data repositories can provide a wide array of well documented data but consist primarily of small data sets. Massive public databases can provide real, complex data but are generally difficult to access. Large databases are important for statistics education because they can provide students with opportunities for complex and engaging statistical exploration. I have developed a point-and-click online interface that allows students to take subsamples from a variety of large databases by specifying random sampling schemes. It is intended for introductory students to be able to treat the databases as a population from which they can obtain random samples. While taking subsamples is not the way a professional statistician would work with big data, the application holds pedagogical value and serves as proof of concept that teaching tools can be built to simplify connection to large data sources. The tool uses `R` as a computation engine, manages the JavaScript interface with the `shiny` package and connects to the database with the `RMySQL` package. The tool can be viewed at shiny.stat.iastate.edu/karstenm/ShinyDatabaseSampler. In the future I will develop a second `shiny`-based application that allows users to generate summary statistics and graphical displays of data from large databases by specifying aggregation options.

In the area of statistical graphics, I am interested in the development of visualizations for big data and in interactive data visualization. I am currently working with development loss functions for binned scatterplots and researching the properties of those loss functions. Binned scatterplots are the two dimensional analog of a histogram that use shading to display counts on a two dimensional binning grid. Binned scatterplot are useful in situations with large data, where traditional scatterplots would suffer from overplotting. However, in the process of binning we inherently lose information about the location of the original data points. We can quantify this visual loss through the distances between the points and the visual centers of the bins. Using increasingly smaller bin sizes will reduce the location information lost in aggregation but will come at the direct cost of computation time; a non-trivial consideration when dealing with massive data sets. The goal of my current research on this topic is to optimize the trade-off between computation time and visual information loss.

The fields of undergraduate statistics education and statistical graphics pose interesting problems and many opportunities for collaborative, interdisciplinary research. My current research has been benefited immensely by working with people with expertise in statistics, human computer interaction, education, experimental design and perceptual psychology. I will carry my passions, ideas and collaborative approach in research into my academic career.