"Science Stays True Here"
Journal of Mathematics and Statistical Science, Volume 2016, 264-273 | Science Signpost Publishing

Open Data at the Interface of Mathematics and Civics Education: Challenges of the Data Revolution for the Statistics Curriculum

Joachim Engel

Ludwigsburg University of Education, Germany

engel@ph-ludwigsburg.de

Abstract

The availability of data of sheer unlimited scope and magnitude changes in radical ways our access to information. Open data nowadays are public domain via National Statistics Offices, UN agencies and NGOs like Gapminder, IPUMS etc. Powerful digital tools for visualizing complex multivariate data sets are available on the web. However, statistics education in schools and colleges lacks behind these developments. This paper examines implications of the data deluge for redesigning curricula that address the needs for active citizenship in the digital age. Besides the mathematics involved to understand the underlying statistics, exploration of data about society also addresses what certain values (equity, fairness, human rights, etc.) mean to students. This discussion can enrich the discourse in the classroom and engages students equally in context, content and statistics.

Keywords: data deluge, statistics education, statistics curriculum, modeling

1. Introduction

A vibrant democracy requires informed citizens who can monitor and discuss societal progress towards economic and social goals and human rights. However, this requires the ability to explore, understand and reason about information of a multivariate nature, as most social problems and trends involve complex data and multiple variables. Sound evidence-based decision-making in private and public life depends upon a certain level of quantitative reasoning skills to understand important social issues. While open data nowadays are easily accessible through National Statistics Organizations, UN offices and NGOs like Gapminder etc., statistics educators face the challenge to teach quantitative skills needed to

understand and interpret these data. Besides strengthening the civil society through empowering people to understand, support and participate in evidence-based decisions in private as well as in public life, integrating issues of monitoring social progress provides a strong experience to students of statistical analyses playing a role in understanding the pressing social and political issues of our time. Acquiring skills in understanding and interpreting large scale multivariate data is an important step in enabling concerned citizens to impact policy decisions and to strengthen civil society.

Statistics nowadays is part of the school curriculum in most countries. Yet statistics education both in high school and colleges is lagging behind the demands for informed citizenship as described above. Relevant datasets from the social and health sciences often have a more complex multivariate structure than the data students work with in mathematics where they 'learn' statistics. Reasoning with large scale multivariate datasets and understanding their display in dynamic graphical representation requires different skills than the analysis of small or moderate sample size uni- or bivariate data that dominate todays curricula. Drawing information from very big multivariate data sets involves statistical principles quite different from dealing with small samples. While inferential methods like hypothesis testing become quite irrelevant when dealing with large samples, working with multivariate data necessitates an awareness of interactions, confounders and non-linear relations. Understanding these data requires an appreciation of statistical modeling, the capacity to employ exploratory techniques and basic knowledge about methods of data collection, together with critical thinking skills.

2. Open Data, Technology and the Curriculum

Over the last decades, NSOs in many countries have increased their efforts to provide data to inform the public in recognition of their central role not only for policy makers and stakeholders in business and economy but also for informing civil society. Government agencies and NGOs in many countries make abundant data as raw material available to the general public to encourage public engagement; recent initiatives such as data.gov in the U.S., data.gov.uk in the UK and govdata.de in Germany refer explicitly to political objectives, in particular the promotion of democratic processes and civil participation by allowing citizens' access to data so as to stimulate debate and to promote decision making. Modern technology shapes the way evidence is used to influence public opinion and policy. With powerful technological advances dynamic interactive visualizations of large multivariate datasets are made accessible, some from non-government organizations like the Gapminder Foundation

(www.gapminder.org) with the promise to enable users to do their own data exploration and hence acquire evidence-based new insights.

Despite a strong boost for statistics in school curricula after recognizing the role of statistics in society in recent years, the actual practice of teaching statistics at schools in most countries is inadequate for preparing young people to be engaged in the civil society. Attention in statistics classes focuses too much on mastery of technique (e.g., regression formulas, analysis of variance) and mathematical underpinnings, and on simple datasets, rather than on the skills required for understanding patterns and changes in socially meaningful phenomena, such as comprehension of complex graphs and tables and understanding underlying causal factors. There is– with the acknowledgment of exceptions of islands of excellence - little attention to statistical thinking, and no exemplification of the power of statistics as an aid to understanding social (or other) phenomena. Today's statistics curriculum is dominated by the statistics developed in the pre-computer age. If a curriculum were to be devised from scratch in the light of current statistical knowledge, it would not be dominated by univariate parametrized distributions and linear bivariate models. It would include bootstrapping, nonlinear regression and multiple regression models, all of which are computationally intensive but conceptually accessible (Engel, 2010; Pfannkuch & Budgett, 2014; Hesterberg 2014).

3. Investigating Economic Discrimination and the Gender Pay Gap

We illustrate statistical topics involved in understanding multivariate social data along an example of investigating economic discrimination. Any serious discussion of the application of statistics in the area of social sciences has to address the issue of reliable and valid measurement and operationalization. In general terms, economic discrimination is usually defined as the difference in average wage rates of minority and majority workers who, reasonably assumed, have equal productive capacities (Cain 1984). The International Covenant on Economic, Social and Cultural Rights, signed and ratified by 160 member states of the UN (and signed but not ratified by 7 additional states), confirms in Article 7 the right for equal wages for work of equal value without distinction of any kind, in particular without discrimination on the basis of gender. In the following we focus on the difference in earnings between male and female employees, known as the gender pay gap.

In 2012, according to the German National Statistics Office (NSO), Germany had one of the largest raw GPG among the European countries, with women earning 22% less than men. Income data from a random selection of 59,504 adults from the 2006 National Income Structure Survey is accessible for

academic purposes from the German Statistical Office's website[1], which allows an accurate analysis of authentic data as well as provides the opportunity to do some specific analyses for sub-populations. A straight and direct computation of averages leads to a mean pay rate of 18.02 € for men versus 14.66 € for women. Hourly pay rates may depend on the total number of working hours, i.e. it may matter if the work is part time or includes extra or odd working hours. For the German income survey data men worked an average of 159 hours per months compared to 135 hours for women. Figure 1 displays the data and regression lines for monthly pay versus total hours worked per month. The slopes (18 for men versus 14.7 for women) confirm the sizeable differences in income. The data cloud, however, suggests that the assumption of a linear relationship between the two variables under consideration and the arithmetic mean as measure of comparison is more than questionable. More robust regression techniques including nonlinear and nonparametric regression may be more appropriate.

Fig. 1. Monthly Income and total working hours for 59,504 German employees from the National Income Structure Survey plus regression lines for men and women.

Figure 2 shows kernel density estimate for gross monthly wages and for hourly pay, for the total population as well as separate for men and women. Notice the skewness which is typical for income distributions. Also, monthly income beyond 7000 € is cut-off explaining the bump at this income value. Because of notable outliers, one might consider the median (16.16 € for men, 13.60 € for women) as a more adequate measure of comparison which reduces the GPG to 15.81%.

[1]http://www.forschungsdatenzentrum.de/campus-file.asp

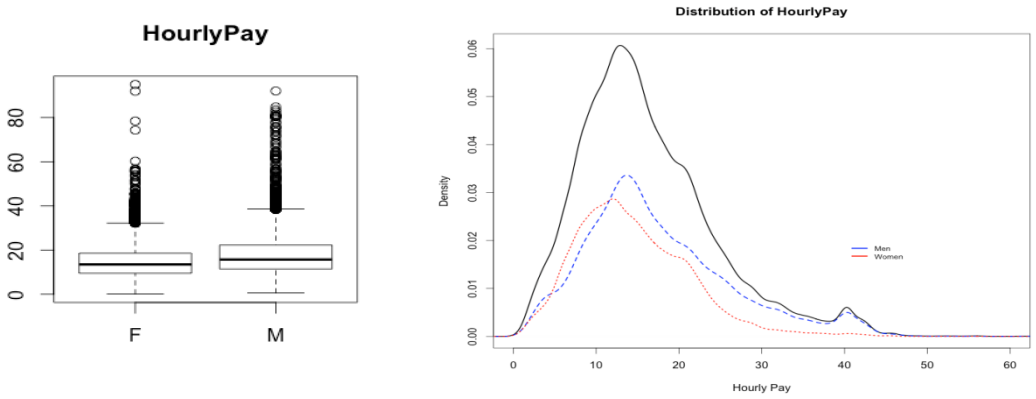

Fig. 2. Distribution of monthly and hourly income from the German income structure survey, dashed line for men, dotted line for women, solid line for men and women together

It is important to differentiate between the *unadjusted* (also known as *raw*) wage gap and the *adjusted* wage gap. The unadjusted or raw pay gap does not take into account differences in personal (e.g., age, education, the number of children, job tenure and occupation) and workplace characteristics (e.g., the economic sector and place of employment) between men and women. Parts of the raw pay gap can be attributed to the fact that women, for instance, tend to engage more often in part-time work and tend to work in lower paid industries. The remaining part of the raw wage gap that cannot be explained by variables that are thought to influence pay is then referred to as the adjusted gender pay gap and is interpreted as being discriminatory. However, interpreting the adjusted gap as being the only discriminatory component may fall short of the reality. Part of the pay gap which is attributed to observed differences in characteristics (such as age, education, hours worked etc.) may still reflect the outcome of discriminatory social processes. A major reason for this gap may be related to the fact that women tend to choose lower-paid professions, or have their jobs valued less favorably. The origins of these factors could be judged as being discriminatory in themselves – that is, when they are rooted in gender stereotypes of male and female occupations.

Keeping only one or a few variables fixed, may even lead to an increased GPG. For example, looking at graduates of Universities of Applied Sciences (in Germany called *Fachhochschulen*) allows one to compare men and women with roughly the same academic degree: a B.A. or B.A. equivalent. Surprisingly, the GPG with women earning 31.3% less than men is particularly large, despite their comparatively equal qualification. A closer look reveals that most male graduates from this type of academic institution work as engineers while many female graduates got their degree in the financially much less attractive field of social work. Obviously, the GPG has nothing to do with equal pay for equal work. Instead, it merely indicates that men generally occupy positions that pay more. A major reason for this gap may be related to the fact that women tend to choose lower-paid professions or have their jobs valued less favorably. In cultures with more traditional family structures (e.g. husband, wife and two children), the husband is considered the main wage earner while the wife's income is seen just as a welcome supplement. The origins of these factors could be judged as being discriminatory in and of themselves – that is, when they are rooted in gender stereotypes of male and female occupations.

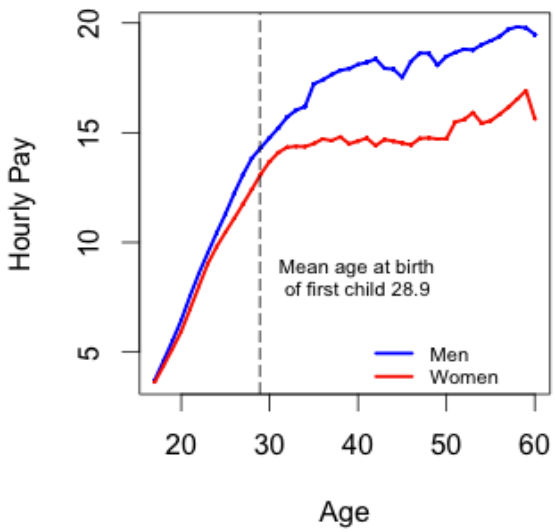

Fig. 3. Hourly Wages for men and women in dependence of age. The dashed line corresponds to the average female age at first child birth

In social science most functional relationships are non-linear despite the noted preference for linear modeling in quantitative social research. A closer look at the GPG in dependence of age reveals another remarkable structure. A nonparametric data smoother (Cleveland and Devlin 1988) shows that men and

women increase their earnings at roughly equal pace during their twenties, but at around age 30 the female income stagnates while the male average income keeps on increasing. This means that the male advantage in earnings is gained primarily during the thirties. Here it is noticeable, that the average age of a German woman at the birth of her first child is at 28.9. However compelling this argument may be, we have to be aware that the data are observational, based on a cross sectional sample and not a longitudinal study. Some of the income discrepancy may also be due to the fact that habits and roles in society are changing very slowly and gender equality may easier be reached in the younger generations. Another noteworthy observation can be made between the accessibility to day care and GPG. In German society, women frequently work part-time while being the primary caregiver for children or the elderly at home. Looking at the GPG in each of the 16 German states in dependence of the proportion of children under age 3 in day care reveals a remarkable relationship with a much lower GPG in states that provide a high percentage of day care allowing young mothers to continue working full-time (see Figure 4). Interpretation of this relationship opens classroom discussion for important statistical topics: the problem of concluding from observed correlation to causation, and the possibility of confounders.

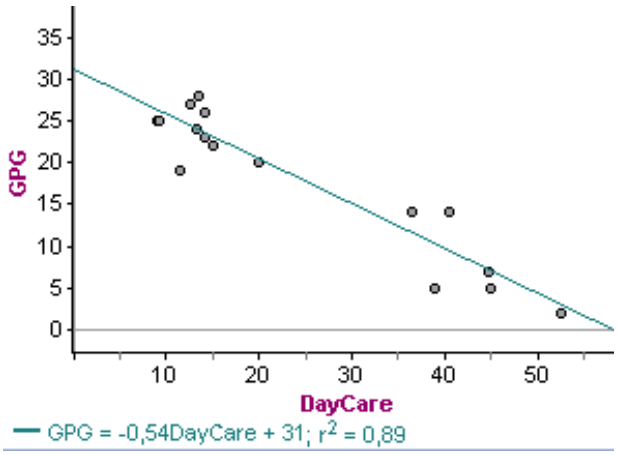

Fig. 4. Average GPG and percentage of children under age 3 in day care in 16 German States

While the observed correlation is striking, the six states with high day care coverage and low GPG are all located in the Eastern part, which formerly belonged to communist East Germany. Are they culturally comparable to states in the western part of Germany regarding how they value employed women?

Figure 5 presents a table of statistical issues involved in analyzing multivariate social data. These topics are mostly neglected in current curricula

1. Exploring the data:
   - qualitative descriptions of the data
   - the story behind the data
2. Operationalizing the data:
   - How were the relevant variables defined?
   - How were the variables measured?
3. Data quality and provenance
   - How were the data collected? Sampling bias?
   - Are the responses truthful and trustworthy?
4. Comparing Distributions
   - Mean, Median, Variance
   - Boxplots
   - Density plots
5. Nonlinearity issues, e.g. age versus pay gap
6. Conditional probability, e.g. investigating subgroups like people with similar educational backgrounds
7. Simpson's paradox
8. Critical Thinking:

Fig. 5. Statistical topics involved in analyzing multivariate data

4. Conclusions

Developments with open data offer unprecedented access to large scale authentic data sets on a huge variety of topics relevant to public policy and personal happiness (Ridgway, Nicholson, McCusker 2013). Including open data on relevant topics into teaching statistics promises to give students a strong experience that statistical analyses matter, are an important tool for evidence based decision making and help to understand the pressing problems of society. This holds in particular when instructing social science students, who traditionally tend to be less interested in formal mathematics.

However, including multivariate complex data sets into teaching implies particular challenges. Successful use of such data requires different skills than more traditional contents of statistics teaching. Key skills involve a critical appreciation of data provenance and quality, and an understanding of statistical ideas associated with multivariate analysis of large data sets. While inferential techniques like hypothesis testing may be less relevant when analyzing very large data sets, important ingredients of multivariate thinking imply the search for interactions, the awareness of confounders, Simpson's Paradox and knowledge about the pros and cons of observational studies and designed experiment. Important

mathematical methods include discovering and modeling functional relationships between two or more variables beyond linear regression including exploring nonlinear relationships either through non-linear modeling or through purely exploratory smoothing techniques.

References

[1]. Cain, G. (1984). The economics of discrimination: Part I. Focus 7 (2), University of Wisconsin-Madison, Institute of Poverty Research.

[2]. Cleveland, W.S. & Devlin, S.J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, Journal of the American Statistical Association, Vol. 83, pp. 596-610.

[3]. Engel, J. (2007). Daten im Mathematikunterricht: Wozu? Welche? Woher? Der Mathematikunterricht, 53 (3), 12-22.

[4]. Engel, J. (2010): On teaching bootstrap confidence intervals. In: C. Reading (Ed.), Proceedings of the Eighth International Conference on Teaching Statistics. Voorburg, The Netherlands: International Statistical Institute.

[5]. Engel, J. (2013). Statistics Education and Human Rights Monitoring. In: S. Forbes and B. Phillips (Eds.) Proceedings of the joint IASE/IAOS Satellite conference Macao, 2013, http://iase-web.org/documents/papers/sat2013/IASE_IAOS_2013_Paper_2.5.1_Engel.pdf

[6]. gapminder http://www.gapminder.org

[7]. Hesterberg, T. (2014). Bootstraping for learning statistics. In: K. Makar; B. de Sousa & R. Gould (Eds.), Proceedings of the Ninth International Conference on Teaching Statistics, Voorburg, The Netherlands: International Statistical Institute.

[8]. Nicholson, J.; Ridgway, J., & McCusker, S. (2013). Getting real data into all curriculum subject areas: Can technology make this a reality? Technology Innovations in Statistics Education, 7 (2), http://escholarship.org/uc/item/7cz2w089

[9]. Pfannkuch, M. & Budgett, S. (2014). Constructing inferential concepts through bootstrap and randomization-test simulations: a case study. In: K. Makar; B. de Sousa & R. Gould (Eds.), Proceedings of the Ninth International Conference on Teaching Statistics,Voorburg, The Netherlands: International Statistical Institute.

[10]. Ridgway, J.; Nicholson, J., & McCusker, S. (2013). Open data and the semantic web require a rethink of Statistics Teaching. Technology Innovations in Statistics Education, 7 (2), http://escholarship.org/uc/item/6gm8p12m

[11]. Ridgway, J. & Smith, A. (2013). Open data, official statistics and statistics education – threats, and

opportunities for collaboration. In: S. Forbes and B. Phillips (Eds.), Proceedings of the joint IASE/IAOS Satellite conference Macao, 2013.

http://iase-web.org/documents/papers/sat2013/IASE_IAOS_2013_Paper_K3_Ridgway_Smith.pdf