OPEN DATA, CIVIL SOCIETY AND MONITORING PROGRESS: CHALLENGES FOR STATISTICS EDUCATION

Joachim Engel
Ludwigsburg University of Education, Germany
engel@ph-ludwigsburg.de

The paper discusses the role of statistical knowledge for active participation in democratic processes. It is based on the assumptions that knowledge and skills to reason adequately with data are an important prerequisite for the functioning of democracy in our mass societies. 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. Drawing information from very big multivariate data sets may involve statistical principles different from dealing with small samples. Besides strengthening the civil society, integrating issues of monitoring social progress lets students experience that statistical analyses play a role in understanding the pressing social and political issues of our time.

INTRODUCTION

Knowledge and skills to understand and analyze trends in data are an important prerequisite for active citizenship in democratic societies. Sound evidence-based decision-making in private as well as in public life of our information-laden world requires a certain level of statistical literacy that implies critical thinking and reflecting on metadata. Monitoring societal progress towards respect of economic, social, and cultural human rights is mainly about analyzing trends in mass phenomena that may contradict the purposes and visions of an open society (Engel 2013). The questions as to whether women are disadvantaged in their careers, immigrants are sufficiently integrated, workers' demands for wage increases are justified, or access to higher education is too strongly determined by the socio-economic background – all have to be judged largely on a quantitative level whether a society keeps up with the promises of equity and fairness to everyone. To assess these trends requires fluency when reasoning with quantitative evidence. Acquiring skills in understanding and interpreting large scale multivariate data is an important step in enabling concerned citizen to impact policy decisions and to strengthen civil society. The ultimate goal of teaching statistics based on authentic open data is empowerment – giving people tools and know-how so they can better make informed decisions in their private and public life. Analyzing these data, however, requires specific skills that may not receive sufficient focus when teaching emphasizes inference from small univariate or bivariate data sets.

While, by many accounts, it is a difficult task to motivate students to learn statistics in a formal and purely theoretical framework, investigating issues related to health, environment and social justice issues provides ample opportunities to apply statistical knowledge and critical thinking to pressing real-world problems, on the basis of authentic data. The benefits and challenges of using authentic data that represent real-world problems for the learning process in contrast to fake or made-up data have been recognized by many statistics teachers (e.g., Engel 2007, Hall 2011, Nicholson, Ridgway & McCusker 2013). However, there are also notable pitfalls and caveats to consider when analyzing huge data sets retrieved from internet sources or surveyed by unknown data collectors. Metadata (Who collected the data? How? In whose interest?) have to be taken into critical scrutiny. Further, any serious discussion of the application of statistics to social studies has to address the issue of reliable and valid measurements and operationalization. Emphasis is needed on the critical evaluation of what appropriate measures or displays are.

TECHNOLOGY, OPEN DATA AND THE MEDIA

Modern technology shapes the way evidence is used to influence public opinion and policy. More recent developments include the following: 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

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. iase-web.org

processes by allowing citizens' access to data so as to stimulate debate and to promote decision-making. 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.

Traditional print media increasingly provide interactive abilities for data exploration on their web pages to support their news report and to allow readers far more in-depth exploration than a traditional newspaper article. With the emergence of 'data-driven journalism,' journalists are making good use of rich data, and create interactive websites that allow in-depth research far beyond what traditional print journalism made possible (Gray, Chambers & Bounegru 2012). An exciting range of visualizations is being developed to give users the potential of exploring large scale multivariate datasets (Ridgway & Smith 2013). The rise of data driven journalism will enhance the use of data in the media, and will exacerbate the problems of interpretation and misinterpretation. The provision of powerful tools alone, without acquiring appropriate skills, does not necessarily lead to empowered citizen (Ridgway, Nicholson, McCusker 2013). Many of these new technological tools are not accessible for lay people or people with moderate quantitative skills. Statistics nowadays is part of the school curriculum in most countries. Yet 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 does the analysis of small or moderate sample size univariate or bivariate data sets that dominate today's curricula. As an example for the potential and the challenges large authentic data sets offer for teaching statistics, we illustrate an analysis of authentic German income data, in particular we compare average pay rates for male and female employees and look at what is known as the gender pay gap.

INVESTIGATING ECONOMIC DISCRIMINATION: THE GENDER PAY GAP

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 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 (http://www.forschungsdatenzentrum.de/campus-file.asp), 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. Figure 2 shows density estimates 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 raw GPG to 15.81%.

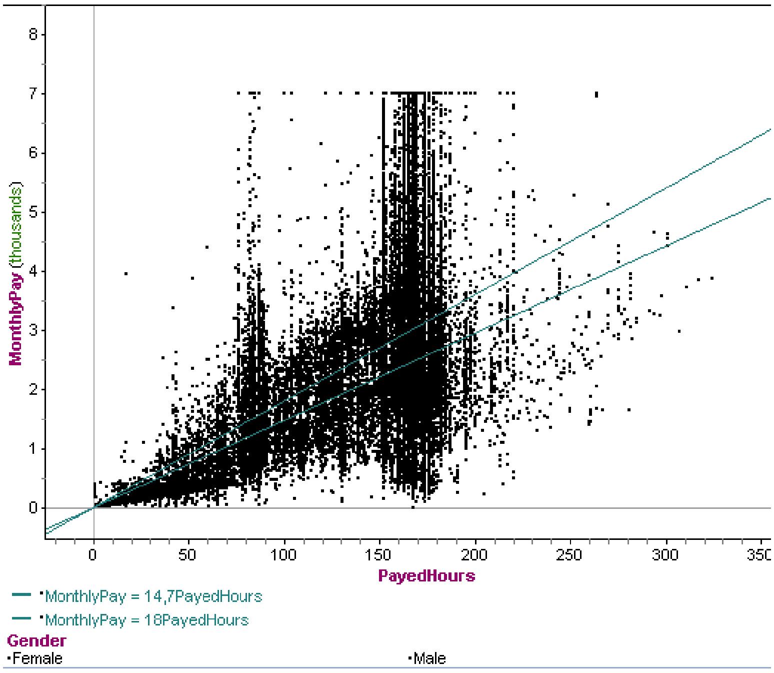

Figure 1: Monthly Income and total working hours for 59,504 German employees from the National Income Structure Survey plus regression lines for men and women.

When investigating the GPG it is important to differentiate between the *unadjusted* (also known as *raw*) wage gap and the *adjusted* wage gap. The unadjusted or raw gender pay gap (GPG) does not take into account differences in personal (e.g., age, education, the number of children, job tenure, position and occupation) and workplace characteristics (e.g., the economic sector and place of employment) between men and women. Part 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 branches. The remaining part of the raw wage gap that cannot be explained by variables thought to influence the pay is then referred to as the adjusted gender pay gap. The adjusted GPG is based on regression with various covariates such as education, position, branch, job experience, etc. For the German 2012 data, the National Statistics Offices computed an adjusted GPG of still 8%.

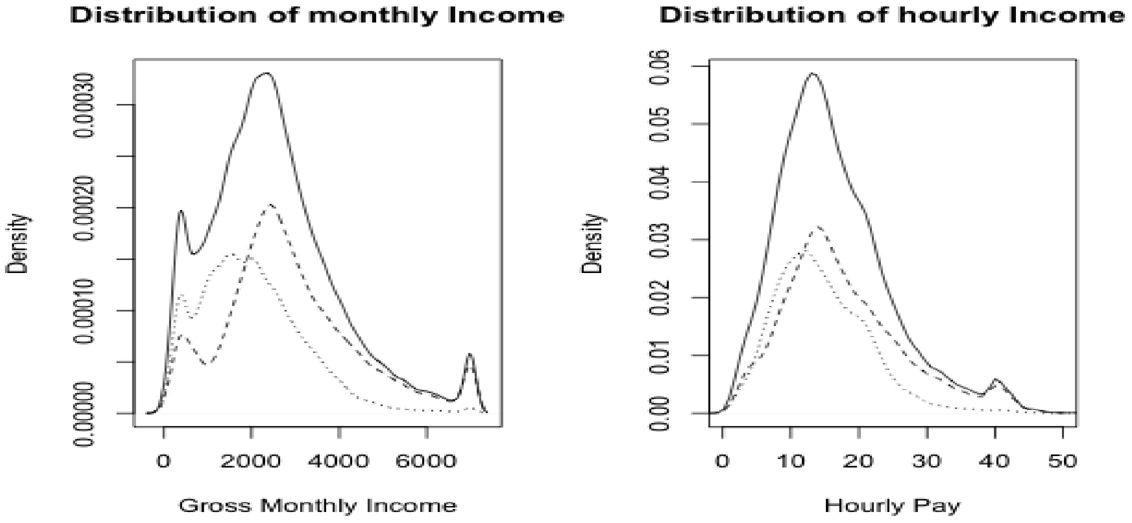

Figure 2: Distribution of monthly and hourly income from the German income structure survey, dashed line for men, dotted line for women, solid line for men and women together

However, even when using the adjusted GPG, how certain can one be to have adjusted for *all* relevant variables except gender? As with any data from observational studies, there is always a possibility for hidden confounding variables that have been overlooked. In fact, some critics of the GPG claim that in most countries, the GPG may completely disappear if *all* relevant variables but gender are taken into consideration.

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.

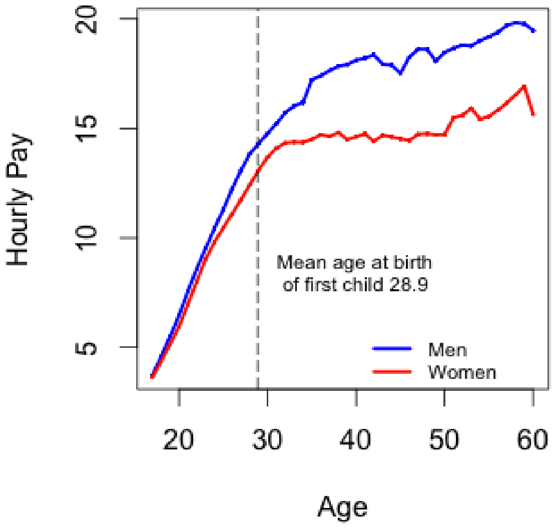

Figure 3: Hourly Wages for men and women in dependence of age. The dashed line corresponds to the average female age at first child birth

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 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 habits and roles in society are changing very slowly and gender equality may easier be reached in the younger generations.

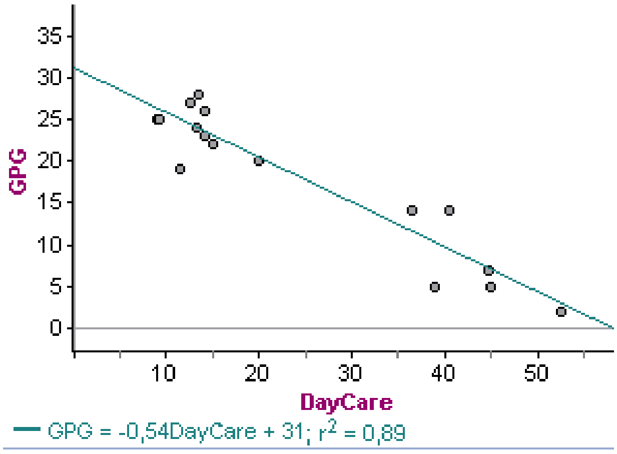

Figure 4: Average GPG and percentage of children under age 3 in day care in 16 German States

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 the age of 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. 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?

CONCLUSION

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 and 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 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

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

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting, *Journal of the American Statistical Association*, *83*, 596-610.

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

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*. http://iase-web.org/documents/papers/sat2013/IASE_IAOS_2013_Paper_2.5.1_Engel.pdf

Gapminder (n.d.). http://www.gapminder.org

Gray, J., Chambers, L., & Bounegru, L. (2012). *The data journalism handbook*. O'Reilly Media. http://datajournalismhandbook.org/

Hall, J. (2011). Engaging teachers and students with real data: Benefits and challenges. In C. Batanero, G. Burrill and C. Reading (Eds.), *Teaching statistics in school mathematics – Challenges for teaching and for teacher education. A joint ICMI/ IASE study: The 18th ICMI Study*. Dordrecht, the Netherlands: Springer.

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

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

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