DATA SCIENCE EDUCATION IN SECONDARY SCHOOL: HOW TO DEVELOP STATISTICAL REASONING WHEN EXPLORING DATA USING CODAP

Lea Budde, Daniel Frischemeier, Rolf Biehler, Yannik Fleischer, Dietrich Gerstenberger, Susanne Podworny & Carsten Schulte
University of Paderborn
dafr@math.upb.de

Data Science has become an emerging field at the intersection of statistics, computer science and application fields and this discipline requires "new skills" to be enabled to explore for example large and messy datasets, so-called Big Data. Because of this emerging relevance we started an interdisciplinary project between statistics and computer science education, which is initiated by Deutsche Telekom Stiftung, with the aim to concretize Data Science and its implications for schools. We offer an innovative and interdisciplinary approach on how to implement Data Science in secondary school under the consideration of the need of "new skills in statistics education". In this paper we will report on an introduction into Data Science at secondary school with the focus on exploring multivariate data with CODAP.

INTRODUCTION

Big masses of data, so called Big Data, are available for many contexts and a competent understanding of data has become more and more important in these days. In a vibrant democracy gaining a robust understanding of data, for example with regard to statistics in the area of migration, health and poverty, so called civic statistics (Engel, Gal, & Ridgway, 2016), becomes even more important. Data Science has become an emerging field, bringing together several disciplines like statistics and computer science and requiring "new skills" to be enabled to explore Big Data (Ridgway, 2016; Gould, 2017). Thus, we started an interdisciplinary project between statistics and computer science education, which is initiated by Deutsche Telekom Stiftung, with the aim to concretize Data Science and its implications for schools. Bringing Data Science into secondary school and developing an early competent understanding of data is highly relevant. We offer an innovative and interdisciplinary approach how to implement Data Science in secondary school under the consideration of the need of "new skills in statistics education". As a first component of the curriculum, we have designed (and already conducted twice) a project course to introduce secondary school students into Data Science in the frame of a Design-Based Research setting (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). The course itself includes three modules: "Data and data detectives" (Data exploration with CODAP and Jupyter Notebooks), "Machine Learning" (Big Data and Machine Learning – Decision trees and artificial neural networks) and "Data Science project" (cooperation with partner from industry and administration). In this paper we will concentrate on the first module "Data and data detectives". Here the students experience all phases of a data analysis cycle like PPDAC (Wild & Pfannkuch, 1999) and use CODAP as a data exploration tool. To generate knowledge and products from data, learners are then introduced to the Python programming language in order to experience their further data moves independently.

NEW SKILLS IN STATISTICS EDUCATION

As mentioned in Ridgway (2016), statistics education faces new challenges from the emerging field of data science and the availability of masses of data. Therefore, the well-known PPDAC cycle (Wild & Pfannkuch, 1999) has to be re-interpreted, because validation is missing, modeling is not explicitly mentioned and the data might be in some cases already there (e.g., it is possible to download open data files via Internet, etc.). In addition to that Data Science also includes activities such as data management, tidying data and algorithmic thinking and the notion of data has to be extended and learners and teachers must be aware of new forms of data like data collected by sensors, data generated by the use of social media, data related to pictures and websites, and transactional data. To handle and to explore large amounts of messy and different data, the use of digital tools is inevitable to analyze and explore large and messy datasets (Biehler, Ben-Zvi, Bakker, & Makar, 2013). However, digital tools not only serve as tools for data analysis, but should also themselves be actively reflected and evaluated. To actively reflect on tools and to evaluate the choice and functionality of the

In: P Arnold (Ed.), *New Skills in the Changing World of Statistics Education*
Proceedings of the Roundtable conference of the International Association for Statistical Education (IASE),
July 2020, Online. Voorborg, The Netherlands. ©2020 ISI/IASE iase-web.org/Conference_Proceedings.php

tool requires a handling that goes beyond the simple use of the tool. Learners should not only experience themselves as consumers of ready-made functions of the tools, but should also actively explore, evaluate and, if necessary, change the tool themselves. However, in order to be able to leave the role of a user, not only the functionality but also the inner structure, i.e. the design of the tools must be examined more closely. Questions like: What is implemented by the tool and how? Are there other possibilities of implementation and how can I use these possibilities? Is another method perhaps even more suitable than the one the tool provides me with? These questions show that the use of tools must therefore also be disclosed as an active phase of reflection. In addition to the actual use, such a phase in the PPDAC cycle and the implementation of the course should therefore also be considered. Thus, the tools are not only used as a means to an end, but also as an independent subject of instruction. The question of which tool to choose is highly relevant and depends on the purpose and the knowledge of the learners. On the one hand educational software exists such as TinkerPlots, Fathom or CODAP which do not require much programming commands and offer learners a landscape of statistical exploration activities but are quite restricted in their capacity to apply more advanced techniques (like e.g. multivariate regression analysis) or limited regarding the size of data. On the other hand there are professional tools like R or Python which allow deep exploration methods and activities but also require the use of a specific programming language (for a short reflection of the use of TinkerPlots, Fathom and R when exploring micro data, see Frischemeier, Biehler, & Engel, 2016). Our primary goal is to develop statistical and computational reasoning of secondary school students in collaboration between computer science education and statistics education and therefore to develop a data science course for secondary school students.

UNIT "DATA DETECTIVES WITH CODAP"

Design ideas of the unit "Data detectives with CODAP"

The idea of the introductory unit "Data detectives with CODAP" was that our students were introduced into first reasoning about data in the frame of a data project using educational software. We decided to use CODAP (see https://codap.concord.org; for a detailed description of CODAP features see: Haldar, Wong, Heller, & Konold, 2018) because this tool allows an easy entrance and a quick start (requiring only a URL) and realizes the exploration of multivariate data. In addition to that we wanted our participants to realize a first data project consisting of the five phases of the PPDAC cycle: Problem, Plan, Data, Analysis and Conclusions. In this frame we wanted to confront our students also with meaningful data for them - therefore we decided to use the JIM study (https://www.mpfs.de/), a representative survey on media and leisure time activities of 12-19 year old students in Germany. This survey includes for example information on the use of social media, messengers, Youtube, games, etc. Due to the fact that only the survey form and the aggregated data but not the micro data of the JIM study are freely available, we decided to use the given JIM survey to collect real data from secondary school students (grades 5-12) in Paderborn (n= 215). The JIM-study includes more than 80 categorical questions (e.g., "How often do you use Whatsapp?" with possible answers like 'daily', 'several times a week', 'once a week', 'two times a month', 'once a month', 'less often', 'never') on the media use of German students. In addition to that, we decided to include questions which tackle numerical variables like "How many apps do you have on your smartphone?" or "How many online accounts do you have?". We collected the data with an online survey at the school of the students of the project course. The survey was advertised with a flyer, and 215 students (not representative) took part in the survey. With its 215 cases the JIM dataset might not be regarded as Big Data as such, but with its large variety of over 80 different variables it serves as an interesting and rich dataset to develop a first competent data handling of our students and to enable them to make deep and diversified explorations. Even if the data do not correspond to the conventional sense of Big Data, they give the learner the opportunity to think about systematic evaluations or about possible analyses and calculations of predictions. Furthermore, the use of this data set does not exclude the possibility of discussing afterwards, based on the findings, what are the characteristics of Big Data in contrast to the already known data. Characteristics such as a much broader range of data types, the handling and necessity of data management and data cleaning, and even algorithmic procedures can be motivated by the rather small data set of the JIM study and then worked out as characteristics of real Big Data. In the introductory unit, we then wanted the learners to become involved in these processes themselves as data explorers rather than only as data analysis consumers. Specifically, we want the students in the unit "Data detectives with CODAP" to…

- explore and analyze a multivariate data set with regard to selected adequate statistical investigative questions,
- use/apply basic terms of descriptive statistics and statistical concepts,
- use and evaluate digital tools like CODAP for their data exploration,
- document and present their data analysis in an adequate form.

The whole project was framed in a data analysis cycle such as the investigative cycle of PPDAC and we want our students to experience all five phases of the cycle: Problem, Plan, Data, Analysis and Conclusions. Data management and tiding processes are only discussed shortly in this unit, but the problem of the selection bias of survey data is supposed to be discussed in the presentation section. One fundamental idea of this project is that our participants are enabled to work on the JIM data project autonomously and on their own. On certain stages we nevertheless prepare prompts to support them in their data exploration process. These prompts are: (1) support when generating adequate statistical investigative questions or analyzing data with CODAP, (2) introduction in statistical exploration activities like using different percentages, conducting group comparisons, (3) setting norms for example about adequate data presentation. In Table 1 we see how the five phases of the PPDAC cycle are allocated to the three sessions (each lasting 135 mins) for the unit data detectives with CODAP.

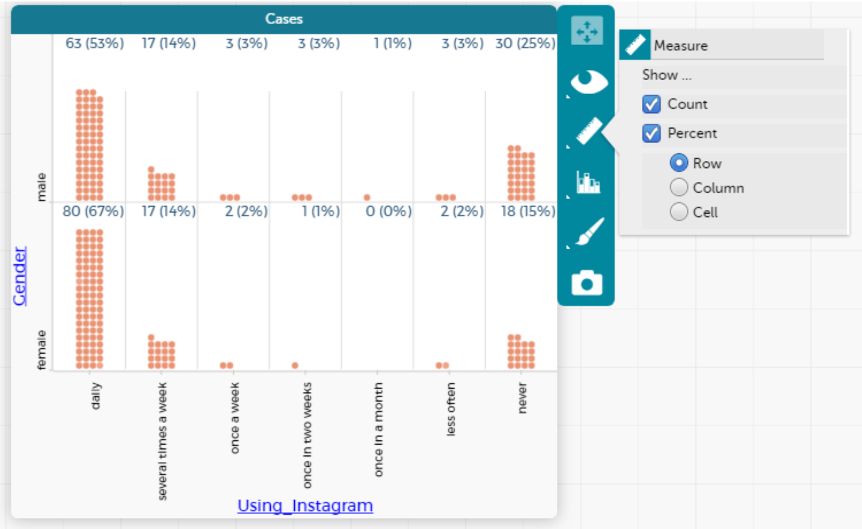
Table 1: Overview on the three sessions in the unit data detectives with CODAP

Session	**Content**	**PPDAC**
1	Introduction in project work, Collection of data, Introduction to basics of descriptive statistics and to the JIM survey, Raising expectations about distributions in the JIM survey, Introduction into CODAP	Problem Plan Data
2	Developing good statistical investigative questions, Exploring the relationship of two categorical variables (using cell, column, row percentages) in CODAP, Comparing groups with CODAP, Setting norms for good statistical representations and presentations, First exploration of JIM data	Data Analysis Conclusions
3	Exploring JIM data, Preparing presentation, Presentation of findings	Analysis Conclusions

Realization of the unit "Data detectives with CODAP"

In the first session of the unit the participants who were secondary school students (17-18 years old) in grade 12 with only few pre-knowledge in descriptive statistics, were introduced into the project, the basics of descriptive statistics and into the survey. For instance, the students were introduced in the intention of the JIM survey. For the project work four different topic areas were presented: (1) Using information media, (2) Using online services, messengers, etc., (3) Using Youtube and (4) Playing games, Computer, Tablet. For the rest of the unit the students were asked to investigate these topics with the given JIM data. The second session started with the generation of statistical investigative questions with regard to these projects. For example, the learners were asked to independently generate statistical investigative questions in relation to the JIM data set. As we know from Arnold (2013) several levels of statistical questions exist and for example Frischemeier and Biehler (2018) have shown that learners (in this case preservice teachers) face massive problems when generating statistical investigative questions. To counteract this, we have developed a Think-Pair-Share-Setting, which enables a systematic generation of adequate statistical investigative questions in one process. Students who first generated yes/no investigative questions like "Do male or female students use more Instagram?" improved their investigative questions after the feedback interventions into an investigative question like "In which way do male and female students differ in their Instagram use?". In the next step the collected data of the JIM Paderborn survey (n=215) was imported in

CODAP. Then it was discussed how CODAP can be used for data analysis. A big difference compared to Excel is that it does not require any specific commands or programming language. We demonstrated our students for example how to (1) display a distribution of a categorical or numerical variable or the relationship between two categorical variables, and how to (2) calculate absolute and relative frequencies or center and spread measures with CODAP. After that, our students were supposed to start a project about their own exploration of the JIM dataset in collaborative learning settings in pairs. Our idea is that the students are enabled to generate statistical investigative questions and to explore the data from a perspective that is meaningful for them. Furthermore, we deepened several statistical concepts like the use of correct percentages when exploring the relationship between categorical variables. Therefore, we provided our students with prompts for the data analysis phase, especially for the handling of different percentages (row, column, cell). In Figure 1 we see an exemplarily screenshot of CODAP when applying row percentages to tackle the investigative question in which way do male and female students tend to use Instagram more regularly.



Measure

Show ...

☑ Count

☑ Percent

● Row

○ Column

○ Cell

Figure 1: Screenshot of comparison visualization between the variables “Using Instagram” and ”Gender” with row percentages in CODAP

As we can see approx. 81% of the female students but only approx. 67% of the male students in the sample use Instagram several times a week or even daily – so one could argue that the female students tend to use Instagram more often than the male students. This kind of comparison can also be made in a more complex way by comparing two categorical variables (both with more than two values), for example when exploring the relationship between the variables “Using Snapchat” and “Using Facebook” (see Figure 2). At this stage our students summarized certain cells and used cell percentages to find relationships between Facebook use and Snapchat use. Six values were thus combined in pairs to form a three-level scale. With this categorization, the students concluded for example that there is a group who uses Facebook and Snapchat “rarely” (29% use Facebook and Snapchat rarely). In a next step we introduced our students how to compare groups with CODAP. Specifically, we put the focus on working out differences between two distributions with regard to center and spread. In the third session the participants were then given 120 minutes of project work to elaborate on their statistical investigative questions generated in session 1.

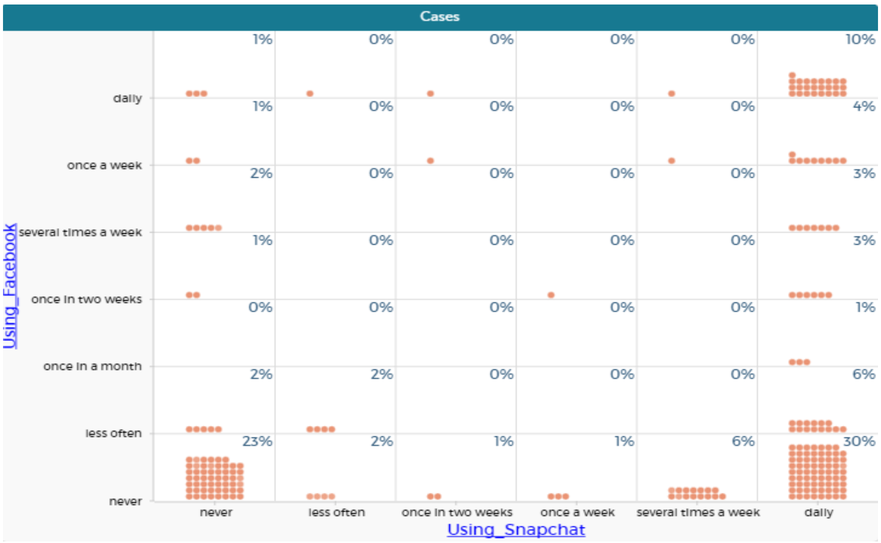
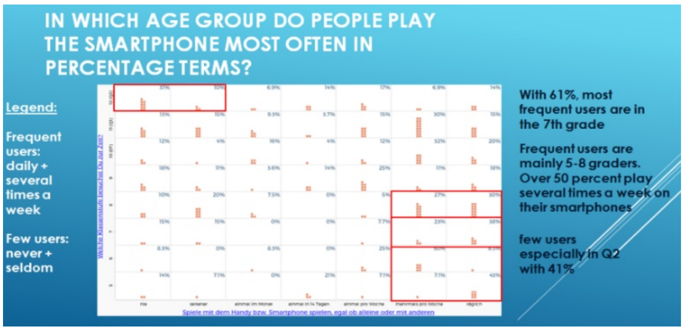


Figure 2: Screenshot of comparison visualization between the variables “Using Snapchat” and ”Using Facebook” with cell percentages in CODAP

The task within their project teams was to explore their statistical investigative question(s), to explore the JIM data in this regard, to document and finally to present the findings in form of a presentation. To get an impression in the work of our students, let us have a look in the exemplary project work of Tim and Marcel who worked on project (4), see Figure 3.



IN WHICH AGE GROUP DO PEOPLE PLAY THE SMARTPHONE MOST OFTEN IN PERCENTAGE TERMS?

Legend:

Frequent users: daily + several times a week

Few users: never + seldom

With 61%, most frequent users are in the 7th grade

Frequent users are mainly 5-8 graders. Over 50 percent play several times a week on their smartphones

few users especially in Q2 with 41%

Figure 3: Slide with findings on the data exploration of the presentation of Tim and Marcel

Tim and Marcel elaborated specifically on the statistical investigative question “In which grades do students spend most time on playing games with their smartphones?”. They used CODAP to create an 8 x 7 contingency table with the variable “Playing on the smartphone” on the x-axis and the variable “grade” on the y-axis. Tim and Marcel defined frequent users and few users (see column left of the CODAP display on the slide) and identified typical groups (framed in red) in the 8 x 7 contingency table, as they have learned e.g. in session 2 of this unit (see Figure 2). Their findings are documented in the column right of the CODAP display.

EVALUATION OF UNIT “DATA DETECTIVES WITH CODAP”

After the described unit we collected feedback from the learners in an online survey. Eight of the 14 students participated in the online survey and rated seven items on a five-point scale from "applies" (5) to "does not apply" (1). As we can see in Table 2, the eight students show in general positive attitudes towards the unit. They consider the unit as motivating and consider that they met the cognitive demands and requirements of the unit. Furthermore, the students are interested in the topic of the JIM study. The relevance for daily life is considered to be important but not so strong compared to the relevance for school. Another important insight is that the eight students consider the handling of CODAP as easy which was one important aspect for our decision. Finally, the students show a positive attitude towards the use of CODAP in their future. All in all, we can say that the participants show positive attitudes towards the unit.

Table 2: Evaluation on the unit data detectives with CODAP

Item	**Mean**
1. The unit "Data detectives with CODAP was motivating". (n=8)	4.25
2. I want to get more information on the topic JIM study. (n=8)	4.00
3. I was able to cope well with the demands of the unit. (n=8)	4.13
4. What I learned in the unit "Data detectives with CODAP" is important for me personally in everyday life. (n=8)	3.50
5. What I learned in the unit "Data detectives with CODAP" should become a general topic in school. (n=8)	4.15
6. The handling of CODAP was easy for me. (n=8)	4.13
7. I would like to work more often with CODAP in the future. (n=8)	3.90

SUMMARY AND OUTLOOK

The unit "Data detectives with CODAP" offers a first starting activity into data science education in secondary school. CODAP serves as a valuable tool for a first data exploration with digital tools and the JIM data represent a valuable dataset which offers meaningful and interesting insights for secondary school students and a plenty of different variables to explore. However, with regard to reach higher visualization capacities and to deal with big and messy data in a next step, the use of programming tools like Python and environments like Jupyter Notebooks become inevitable. An introduction into data science with Python and Jupyter Notebooks is realized after the unit "Data detectives with CODAP".

REFERENCES

Arnold, P. M. (2013). Statistical Investigative Questions - an Enquiry into Posing and Answering Investigative Questions from Existing Data. (Doctor of Philosophy), The University of Auckland. Retrieved from *https://researchspace.auckland.ac.nz/handle/2292/21305*

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for Enhancing Statistical Reasoning at the School Level. In M. A. Clements, A. J. Bishop, C. Keitel-Kreidt, J. Kilpatrick, & F. K.-S. Leung (Eds.), *Third International Handbook of Mathematics Education* (pp. 643-689). New York: Springer Science + Business Media.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design Experiments in Educational Research. *Educational Researcher, 32*(1), 9-13.

Engel, J., Gal, I., & Ridgway, J. (2016). Mathematical Literacy and Citizen Engagement: The Role of Civic Statistics. Paper presented at the 13th International Congress on Mathematical Education, Hamburg.

Frischemeier, D. & Biehler, R. (2018). Stepwise development of statistical literacy and thinking in a statistics course for elementary preservice teachers. In: T. Dooley, & G. Gueudet (Eds.), *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education* (pp. 756-763). Dublin, Ireland: DCU Institute and ERME.

Frischemeier, D., Biehler, R. & Engel, J. (2016). Competencies and dispositions for exploring micro data with digital tools. In: J. Engel (Ed.), Promoting understanding of statistics about society. *Proceedings of the Roundtable Conference of the International Association of Statistics Education* (IASE), July 2016, Berlin, Germany.

Gould, R. (2017). Data Literacy Is Statistical Literacy. *Statistics Education Research Journal, 16*(1), 22-25.

Haldar, L. C., Wong, N., Heller, J. I., & Konold, C. (2018). Students Making Sense of Multi-Level Data. *Technology Innovations in Statistics Education, 11*(1).

Ridgway, J. (2016). Implications of the Data Revolution for Statistics Education. *International Statistical Review, 84*(3), 528-549. doi:10.1111/insr.12110

Wild, C. J., & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review, 67*(3), 223-248. doi:10.1111/j.1751-5823.1999.tb00442.x