Co-funded by the
Erasmus+ Programme
of the European Union

**ProCivicStat ©**

# Syllabus for Civic Statistics

Joachim Engel | Achim Schiller
engel@ph-ludwigsburg.de | schiller01@ph-ludwigsburg.de
Ludwigsburg University of Education, Germany

This syllabus is based on the facets of statistical literacy that make up Civic Statistics (Gal, Nicholson & Ridgway, 2017)[1]. There, facets are described in three groups: (1) **Engagement & Action**, (2) **Knowledge Base** (3) **Enabling Processes**. *Engagement & Action* (facets 1 -3, Meaning for society and policy; Critical evaluation and reflection; and Dispositions) describes features that pertain to any involvement with data about society. Similarly to process standards, these facets cannot be addressed in isolation but permeate any Civic Statistics topic and extend across any knowledge based content. Likewise, the three facets subsumed under *Enabling Processes* (facets 8-11, ICT and search; Quantitative core; and Literacy and communication) are not stand-alone features to be addressed and limited to a specific teaching unit. '*Quantitative core'* is a prerequisite encompassing basic numeracy skills, while the ability to understand, e.g. non-continuous texts, to communicate results and reason about insights from statistical analyses is a general competency embracing any involvement with Civic Statistics. Engagement with data about society most often implies searching for information, probing the credibility of information and retrieving, cleaning and structuring data from the web followed by applying suitable digital tools for visualization, computation and analysis. Throughout a course of Civic Statistics, skills to engage with ICT-based technology such as statistics packages, data retrieval and information search are indispensable tools of the trade. Growing proficiency will be accomplished while being challenged by various tasks throughout the course.
The following syllabus follows a structure that is determined by the ***knowledge base*** (facets 4 to 7) pertaining to Civic Statistics. As its target group the proposed syllabus addresses courses for college or university students, with some basic mathematical skills (e.g., firm grip on school algebra, percentages …) and some elementary knowledge of descriptive statistics. Previous courses in probability and statistics such as *Statistical Literacy for non-quantitative majors* or *Statistical Inference* are certainly helpful, but not a requirement. Elements marked by a star may be optional, depending on the particular target group and the prerequisites of the participating students.

[1] Gal, Nicholson & Ridgway (2017). Facets of Civic Statistics: A conceptual framework

[1] This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Co-funded by the Erasmus+ Programme of the European Union

| Topics | Competences | Why is this topic important for Civic statistics? Which of the knowledge-facets are addressed? | What's new, core elements, some general sample questions and recommended data sets |
|---|---|---|---|
| 1. Basics:<br>• Population and Sample<br>• Attributes<br>• Scales | Knowing about the differences and characteristics of a sample in relation to its population;<br>Knowing about the characteristics of attributes, measuring and scales and the criteria of testing and measuring, knowing about the PPDAC cycle | Basic to any statistical investigation<br>(4) variability, samples, populations and representativeness;<br>(6) sampling and randomization | Using rich multivariate data sets, essential technical terms can be introduced based on a very few (or even a single) data set;<br>Motivation through socially relevant topics. Some attributes may be explored as homework or in-class assignments<br>- *What are the differences between the scales?*<br>- *What makes a sample representative?*<br>***Dataset: Students Alcohol Consumption*** |
| 2. Types of data collections<br>• Survey<br>• Experiment<br>• Observational Studies<br>• Administrative data | Knowing about different types of data collection, their specific advantages and pitfalls; basic knowledge about sampling (simple random sampling, systematic sampling*, stratified sampling* may be helpful, for a more advanced level) | For a suitable interpretation and critical assessment of the results of a statistical analysis it is important to know how data are collected.<br>An understanding of ethical issues associated with the production of data and the use of various research methods is an essential component of Civic Statistics. | Ignored in most "traditional" statistics courses<br>Raising issues of confidentiality and protection of the identity of citizens<br>Emphasize and relate to social context<br>- What are the pros and cons of the respective methods of data collection? |

| | | | |
|---|---|---|---|
| • Opportunistic data | Knowing about bias and variability in any data collection process | (6) Survey types, sampling methods, with attention paid to bias and variability; measurement (reliability and validity); questionnaire design*; web scraping*.<br>(7) survey design* (and associated problems such as non-response or respondent bias) | - *How could you reduce bias in a telephone survey?*<br>- Design a plan for data collection on a topic like, e.g. *"Does watching violent movies encourage aggressive behavior in juveniles?"* |
| 3. Operationalization of variables<br>• Identifying relevant variables<br>• Measuring variables<br>• Different approaches for the same concept (e.g., poverty) | Awareness that the definition of many variables is relatively open (e.g., poverty) and may differ from day-to-day use. The measuring instruments influences the scales of the variables.<br>Knowing about quality criteria for measurements: objectivity, reliability, validity | How variables are operationalized influences the results and thus the possible valuation and interpretation, especially when the meaning of variables in studies is different from everyday use.<br>(5) An essential component of Civic Statistics is an awareness that qualitatively different models may be used to model the same complex social phenomenon.<br>(7) measurement issues (objectivity, reliability and validity; metadata definitions) | ***Dataset:***<br>***Malnutrition UNICEF (focus on African countries)***<br>***vs. Malnutrition HANCI***<br>- *How are the variables defined?*<br>- *How accurate is the measuring instrument?*<br>- *What is the difference between a latent construct and a manifest construct?*<br>- *Multiple choice or open question*? |
| 4. Univariate data analysis<br>• Mean, mode, median | Knowing the formal definitions and the characteristics of the respective parameters as well as | Often data-communicators try to summarize complex phenomena by conveying one or two simple numerical | When using a rich multivariate data set, different parameters and statistics |

| | | | |
|---|---|---|---|
| • Quantiles, trimmed means*<br>• Dispersion parameters: IQR, range, variance and standard deviation<br>• Various distributional shapes (skewness, kurtosis, uni- or multimodality) | their impact and meaning for possible conclusions | representations like mean or the median. Or they try to increase their message with a statement with numerical representations; role of sample size<br>(4) variability; describing and comparing distributions; | can be introduced and explained with only one data set<br><br>***Dataset: Students Alcohol Consumption***<br>- *What are the properties of the mean, mode, median etc.?*<br>- *"Youngsters drink too much alcohol" To argue against this claim would you rather reason with the mean alcohol consumption of young people or with the median? How would a concerned parent argue?* |
| 5. (Graphical) representation of univariate data<br>• Bar plots, bar chart, pie charts<br>• Info graphics<br>• Boxplots<br>• Histograms<br>• Line graphs<br>• Density plots*<br><br>Novel visualizations | Knowing which scale level and graphical representation is appropriate for the given data; how charts and graphs are constructed; knowing about possible manipulations in graphical display as well as some specifics of individual representation. | Knowing about the pros and cons of various graphical representations, misrepresentations and misleading manipulation<br>(4) describing and comparing distributions<br>(5) Representation is a core skill in understanding phenomena. Civic Statistics requires familiarity with sophisticated representations including those that are dynamic and facilitate interactions | A rich multivariate dataset allows you to create many different representations<br>A multitude of graphs, based on the same data set, may be created, thus illustrating the possibility of representing different variables resulting in quite different insights.<br><br>***Dataset: Air Quality dataset*** |

| | | | |
|---|---|---|---|
| | | | - *What must be considered in the evaluation of the respective graphs?*<br>- *Which features in the data can be highlighted with what type of graph? Which features are disguised in which graph?* |
| 6. Bivariate data analysis<br><br>• Correlations, (linear) regression<br>• Residual analysis<br>• Curve fitting*<br>• Comparing distributions, e.g. through boxplots<br>• Understanding 2 x 2 tables<br>• Spurious correlation; correlation caused by confounders;<br>• Different measures of correlations and their pitfalls* (Pearson correlation; Spearman's rank correlation) | Know that correlation does not imply causation.<br>Knowing the formal definitions and the characteristics of the respective concepts as well as which conclusions may be drawn including misconceptions. | In the media correlations are often communicated as causal relationships.<br>(4) variability; describing and comparing distributions; association and correlation; regression; non-linearity; signal and noise; interaction,<br>(7) techniques such as moving averages, case weighting, data smoothing; | Given a rich multivariate data set, there are usually many possible correlations and associations to calculate and to consider.<br><br>***Dataset: Happiness 2015***<br>- What does correlation "only" mean?<br>- What is important in distribution comparisons?<br>- What do regularities indicate in the residual analysis?<br>- Can correlations between differently scaled variables be meaningfully interpreted? |

| | | | |
|---|---|---|---|
| 7. Graphical representation of bivariate data<br>• Scatter plots<br>• Time series, line plots<br>• Age pyramid<br>• (linear) regression<br>• Data smoothers*<br>• Dependent variable: continuous, independent: categorical<br>• Both variables: categorical<br>• Residuals | Knowing how to combine differently scaled variables, knowledge about the meaning and suitability of various graphical representations and the information they convey, including their limitations. | Relations between two metric variables are typically represented in a scatterplot graph; knowing about the pros and cons of different graphical representations and various modeling techniques<br>(4) association and correlation<br>(5) Representation is a core skill in understanding phenomena. Civic Statistics requires familiarity with sophisticated representations including those that are dynamic and facilitate interaction | ***Dataset for different scales: Students Alcohol Consumption***<br><br>***For time series: Swedish crime***<br><br>- What must be considered in the evaluation of the respective graphs?<br>- What distinguishes the graphs from each other and what are the respective advantages and disadvantages?<br>- What could be useful? |
| 8. Multivariate data analysis<br><br>• Understanding the influence of third variables<br>• Simpson's paradox<br>• Understanding higher-order tables*<br>• (Linear) multivariate regression, analysis of variance* | Knowing the formal definitions and the characteristics of the respective concepts; knowing which conclusions may be drawn, including misconceptions | Understanding multivariate phenomena is a core competence for Civic Statistics<br>(4) variability; describing and comparing distributions; association and correlation; regression; non-linearity; signal and noise; interaction<br>(7) synthetic methods where data gathered by conventional survey data is combined with Big Data (e.g. data on mobile phone traffic or web search data) | ***Dataset: Happiness 2016***<br><br>- *Simpson's Paradox*<br>- *What does an explanatory third variable mean?*<br>- *When to expect an explanatory third variable?* |

| | | | |
|---|---|---|---|
| | | | - *Fallacies of confusing correlation with causation* |
| 9. Graphical representation of multivariate data<br>• Comparing distributions<br>• Matrix plots*<br>• Trees*<br>• Times series plot, e.g. Gapminder<br>• Innovative graphical representations, Smart Centre Durham<br>• Geographic graphs | Knowing how to combine differently scaled variables, knowledge about the meaning and suitability of various graphical representations and the information they convey, including their limitations. | There has been a recent explosion in the use of ICT-based representations - Civic Statistics requires the ability to understand and critique novel representations<br>(4) regression; non-linearity<br>(5) Representation is a core skill in understanding phenomena. Civic Statistics requires familiarity with sophisticated representations including those that are dynamic and facilitate interaction | ***Dataset for different scales: Students Alcohol Consumption***<br><br>***For time series: Swedish crime***<br><br>- What must be considered in the evaluation of the respective graphs?<br>- What distinguishes the graphs from each other and what are the respective advantages and disadvantages?<br><br>What could be useful / obstructive contextual knowledge for time series? |
| 10. Indices<br>• Stock or Price Indices<br>• HDI<br>• Gini-Index*<br>• Lorenz-curve* | Critical appreciation of indices as composite measures intended to map complex phenomena on a one-dimensional scale | Several economic or social trends and their effects are presented and explained with indices.<br>(7) definition and importance of indices | ***Dataset: Malnutrition HANCI;***<br>***Measures of social/ economic inequality***<br><br>***Ecological footprint and biocapacity***<br>- Is the definition of the index comprehensible? |

Co-funded by the Erasmus+ Programme of the European Union

| | | | |
|---|---|---|---|
| | | | - Evaluate the significance of the index.<br>- Would you define the index differently? |
| 11. Probability<br>• Variability<br>• Conditional Probability<br>• Risk<br>• Bayesian Reasoning*<br>• Distributions*<br><br>At a more advanced level<br>• Confidence intervals*<br>• Significance tests* | Knowing about basic concepts of probability; understanding the relationship between sample size and variability; understanding the concept of sampling and conditional probability;<br>knowing about standard distributions (discrete and continuous)<br>Knowing how to calculate and interpret a confidence interval as well as knowing how to perform a significance test and to correctly interpret the result. | The concepts of variability, sample and distribution are important for Civic Statistics as foundations..<br>(4) Bayesian inference; bounded estimates; and effect size. Understanding risk is an essential ingredient of Civic Statistics. It relies on probability and conditional probability (including Bayes' theorem), expected values, and subjective utility. | Interpretation of variability, risk, and conditional frequencies in the context of multivariate social data. Basics of probability are needed to avoid misconceptions in interpreting complex multivariate (civic) data (e.g., prosecutors fallacy). |