Developing web-based tools for the teaching of statistics: Our Wikis and the German Wikipedia

Klinke, Sigbert
Humboldt-Universität zu Berlin, Ladislaus von Bortkiewicz Chair of Statistics
Unter den Linden 6
10099 Berlin, Germany
E-mail: sigbert@wiwi.hu-berlin.de

When we started the development of our CD to support the teaching of our basic statistic courses our work was twofold: we had to develop the contents *and* the necessary software ourselves. Maintaining our software development in a university environment turned out to be especially difficult since it is time-consuming. At the same time the available open source software has become much more powerful such that our own development has become superfluous. The use of "standard software" (wikis) allows us to concentrate more on content rather than on software development.

In several projects we have used wikis to support the teaching of our students and publis their projects and homework. However the introduction of the bachelor/master system has reduced the willingness of students to make contributions to our wikis. As a consequence we have now started to involve ourselves and our students in more direct contributions to the German Wikipedia.

1999: MM-Stat I

In 1998 we started to create teaching material for the undergraduate courses "Statistics I" and "Statistics II". A CD was produced on the basis of HTML and JavaScript (Rönz, Müller, Ziegenhagen, 2000). It was structured like a book with chapters and sections using the hypertext capabilities of HTML. Each section consists of one web page with several subpages complementing it with different types of examples (see Figure 1 left) consisting of one web page, too:

- *Interactive examples* allow the students to work interactively with real or simulated datasets.
- *Explained examples* require the understanding of the current section to understand the example, whereas
- *Enhanced examples* also make use of knowledge from previous sections.
- *Information* adds additional information, for example historical information.

Each chapter is finalized with Multiple-Choice-Exercises and the whole CD with a glossary of statistical terms. A lot of JavaScript programming was necessary to run the CD smoothly. With some adaptions the CD will run on modern versions of Internet Explorer. However, maintaining the necessary JavaScript code turned out to be time-consuming. We finally decided to transfer the content to a wiki.

2001: Electronic books

Since most of our research reports and books are produced in LaTeX we decided to create electronic versions from the LaTeX sources. The idea was to generate new electronic versions whenever the book contents would change. We developed a script software, MD*Book, based on `latex2html` to generate HTML (Witzel, Klinke 2002). Figure 1 (middle) shows a section from the book "Applied Multivariate Statistical Analysis" with graphics, formulas and links to interactive programs. Again each section of a LaTeX document is translated into one web page and an additional navigation bar is created at the left. Several books have been produced with it and are freely available on the web.

Late versions of the translations of MM-Stat were also produced with the MD*Book script. The

advantage of using `latex2html` as a basis was the automatic translation of mathematical formulas into embedded graphics. The MD*Book development was discontinued because adaptation of the script to new versions of `latex2html` became time-consuming, too. Moreover, the developers of `latex2html` extended their own functionality to features we covered with MD*Book.

2006: Statwiki

With the Bologna process in Europe the higher education system consists of three cycles which are finalized by awarding Bachelor, Master and PhD degree. The transition from bachelor to master leads to increased mobility of students. Therefore in the second cycle (masters degree) we have a discrepancy in knowledge since students come from different universities. As a consequence we set up a wiki as a dictionary for students and teachers as a joint project by two other chairs and the Ladislaus von Bortkiewicz Chair of Statistics at Humboldt-Universität zu Berlin. In contrast to the Wikipedia, the entries aim more at a reactivation of already learned knowledge.

Different to the MM-Stat CD and the electronic books each statistical term is now in one web page, as in the Wikipedia. As a basis we first used the LatexWiki extension with the Zope Content Management system used at Humboldt-Universität zu Berlin. Since the future development of the LatexWiki was unclear we decided to switch from the LatexWiki to the Mediawiki software which is the basis of the Wikipedia.

With use of the LatexWiki and the *math* extension of Mediawiki, mathematical formulas were easily generated and embedded. For the integration of Figures and Tables into the Latexwiki and the Mediawiki we developed the R extension. It allows us to embbed R programs which generate graphics and tables through R programs (Klinke, Zlatkin-Troitschanskaia 2007).

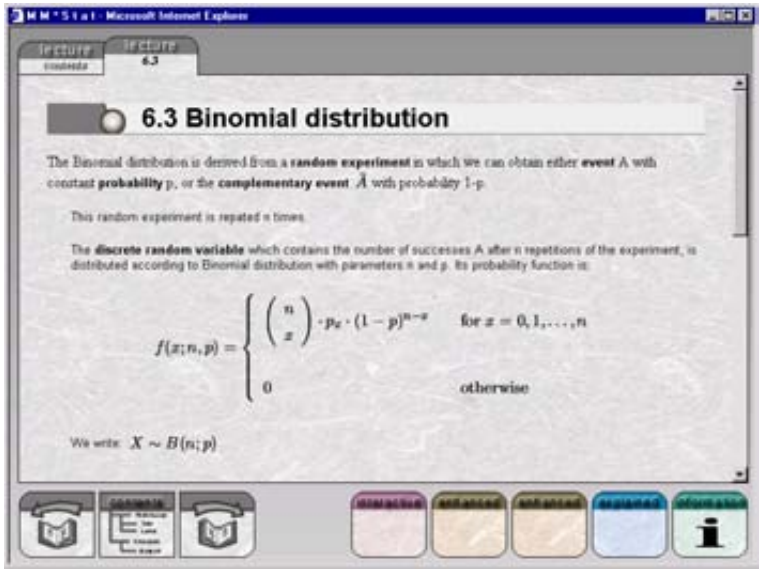
2006: Teachwiki

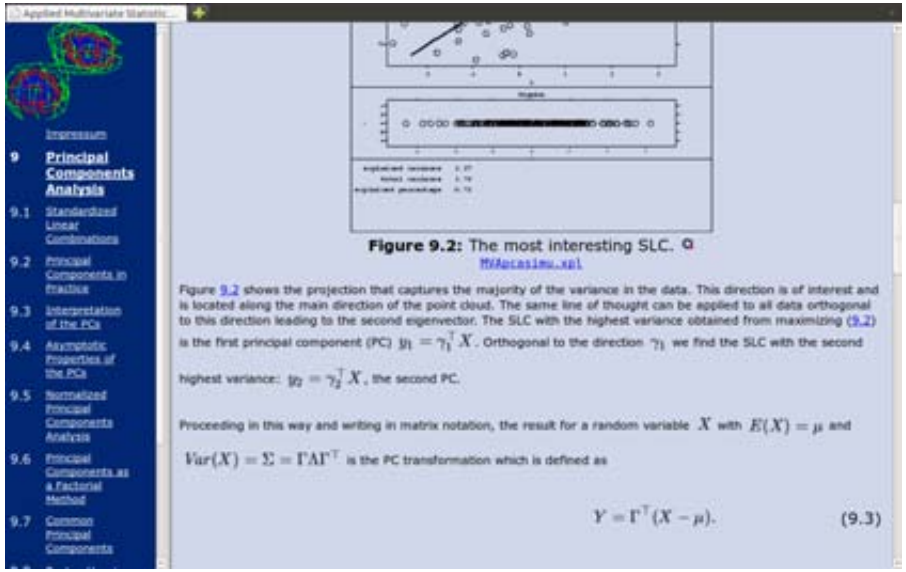
In the same year we set up another wiki for student projects and homework. Usually the papers handed in vanish in a cupboard after grading, and after the compulsory period of record-keeping they are disposed of. However, we think that new students should have the possibility to learn from previous work. We therefore offer students the possibility to create a wiki page with their work, besides the traditional way of delivering their work in paper or electronically as PDF file.

We started with the seminar "Numerical Introductory Course": covering, for example, topics like optimization, random number generation and numerical computation quality in statistical software. The contributions from the students of this course later became part of the wikibook in Statistics (chapter 12: Numerical methods). However, the rules of the wikibook community make it clear that the transfer of complete course work is not desired, we therefore now keep them in our wiki. Later contributions by some students have become the highest ranked German sources in Google for specific queries.

The introduction of the bachelor and master system has drastically reduced the willingness of students to make contributions to the Teachwiki instead of paper-based or PDF-based work. Although a new set up of the Teachwiki offers the possibility to integrate graphics from Wikimedia Commons and the German Wikipedia to facilitate student work, in the last semester no student took the opportunity of putting their work in the wiki. This development encourages us to invest more time in direct contribution/improvement to the German Wikipedia in the area of statistics.

Figure 1: MM*Stat CD example page (left), Electronic book page (middle), Category tree for "Statistik" (right)





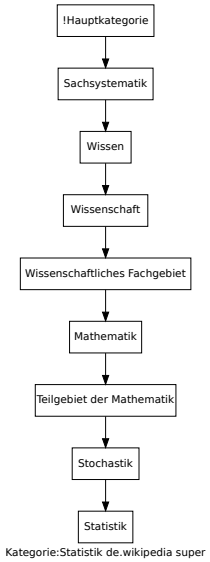


Figure 2: MM-Stat Wiki (left), Link cloud for the lemma "Stichprobe" (right)

2008: MM-Stat II

In 2008 we started to save the contents of the earlier MM-Stat CD into a new wiki available under `www.mm-stat.org` (see Figure 2 left). As with the earlier CD, the wiki is used heavily by students on the basic statistic courses for preparation of exams.

A web page from the MM-Stat CD including all complementing examples has been integrated to one web page into the wiki. Mediawiki extensions ensure a similar navigation as on the CD and the integration of multiple-choice exercises. Only the interactive examples could not be transferred, since they are written in the XploRe language (see Härdle, Klinke, Müller 2000). However, the R extension of the Mediawiki offers some limited interaction possibilities which would be sufficient for the interactive examples. Only the interfaces and the programs need to be translated from XploRe to Mediawiki and R.

The wiki has been expanded by two more books: a book for the lecture "Applied Quantitative Methods" about analysis of questionnaire data which also includes several videos explaining statistical methods. The second book was produced in the framework of an introductory course for R held at the department of educational science at Humboldt-Universität zu Berlin. For more details see Klinke, Kuhlee, Theel, Wagner and Westermeier (2009).

Two further books are in progress: a book for the course "Computer aided statistics" and an exercise collection used in basic statistics courses.

2009: German Wikipedia

The problems with the Statwiki and the Teachwikis leads us to consider whether to immediately contribute to the German Wikipedia in the field of statistics. The German Wikipedia was considered because of two reasons: Firstly, our lectures for bachelor studies are usually in German, which is our main concern, and also because the master programs run partly in German and in English. Secondly, a lot of people contribute to the English Wikipedia such that the area of statistics in the English Wikipedia is already much better developed. We have several aims: improving the quality of existing articles, creating new articles and improving access to the "surrounding" of an article. When we announced our project in the German Wikipedia it was mentioned in the Wikipedia news (see Wikipedia Kurier, 2009).

Category system. Besides the web pages for each term, called "lemma", a hierarchical category system exists in all Wikipedias; see, for example, Figure 1 (right) the hierarchical category tree for "Statistik". Each lemma can be part of one or more categories. All lemmas in one category should belong to one topic. In the category "Statistik" we had more than 500 lemmas and only very few subcategories. As the category tree in Figure 1 (right) shows statistics is part of mathematics. A lot of lemmas are written by mathematicians aiming more at random variables rather than the empirical counterparts and there are also criticisms that they are too mathematical. We have reorganized and extended the category system under "Statistik"; for an excerpt of the current state see Figure 3. The reorganization of the categories suffers from the fact that no common subject classification system for statistics, like the 2010 Mathematics Subject Classification (MCS) system from the American Mathematical Society, is available. The categorization of "62-XX Statistics" in the MCS would often contradict the rule in Wikipedia that a category should have at least 10 lemmas inside to be useful.

Improving quality. From the more than 1000 lemmas in the category "Statistik" and its subcategories we have touched on around 100, mainly related to topics for basic statistics. We have used five categories for the problems we found and corrected. The numbers in brackets give roughly the

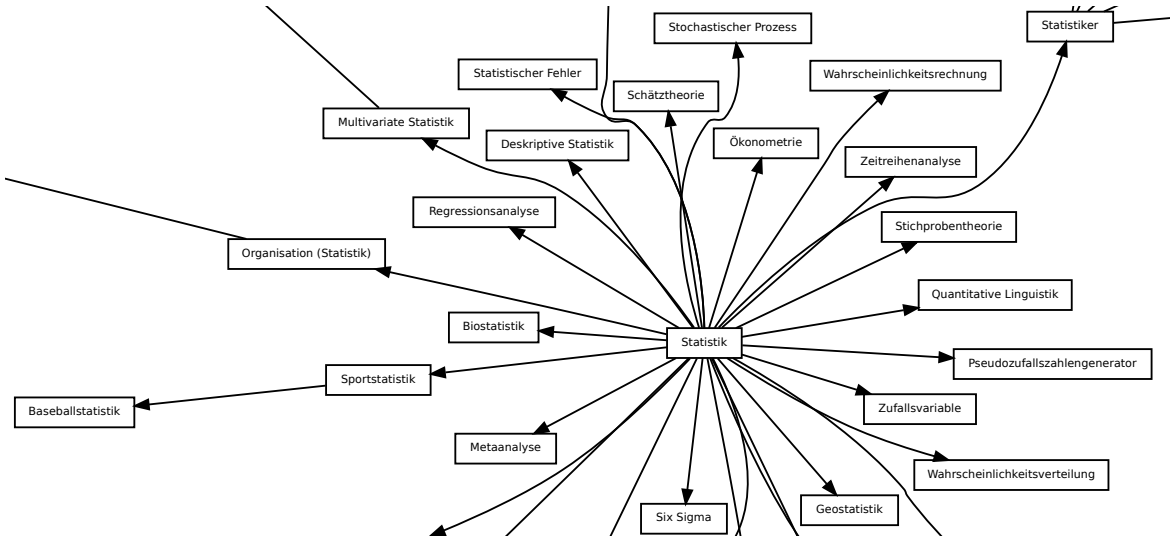
percentage of lemmas which contained this type of problem.
• **Contents** (55%): Modification of mistakes and extension of contents and creation of new lemmas, for example "rank".
• **Cosmetics** (45%): Typos and grammar correction etc.
• **Structure** (35%): Reordering the lemma contents to increase understanding and going from the general to more specific aspects of the lemma. The introductory part of a lemma should be understood by a non-statistician (the grandmother principle: your grandmother should understand what the lemma is about).
• **Reliability** (35%): Adding references, correction of inexact formulations and so on. For example, in the "coefficient of determination" lemma, it was never mentioned in the introduction that the variance of the dependent variable is considered; only the word "explained variance" was used.
• **Access** (35%): New links between lemmas, new redirects to existing lemmas, since the same statistical term is named differently in different application areas.

Improving access. Based on the links between lemmas we have used standard statistical techniques, e.g. multidimensional scaling (distance between two linked lemmas is always one), to visualize the "surroundings" of a lemma. We have generated with R uni- and bivariate link clouds to statistical lemmas; Figure 2 (right) shows the lemma "Stichprobe" (sample). Different colours of the lemma names give information about the link type: black - the current lemma, blue - the lemmas link mutual (bidirectional link), red - current lemma links to lemma, however the lemma does not link back (outbound link) and green - the lemma links to the current lemma, but the current lemma does not link back (inbound link). Since these kind of link overviews are not permitted by the German Wikipedia we have created a subset of lemmas with our link clouds integrated.

Conclusion

Giving up our own software tools and using the wiki software as a basis gave us the chance to concentrate more on content rather than software development. Terminating the work on our own wikis allows us to involve ourselves in the German Wikipedia. However, the quality of the Wikipedia lemmas in statistics is very mixed and partly requires considerably polishing. However, we believe it is not a hopeless task, especially considering that publishers are starting to setup their own online dictionaries in statistics.

Figure 3: Excerpt of current subcategories of the category "Statistik"



Acknowledgements

We would like to acknowledge the support by Humboldt-Universität zu Berlin for our work, namely by the Multimedia-Förderprogramm of the university. Without the initial support of them the MM-Stat CD would have not been possible. The subsequent support in the years 2003-2009 made further work possible. We would like especially to thank our (former) student assistants Sarah Asmah, Paul Giradet, Patrick Lehmann, Vinh Hanh Lieu, Leonie Schlittgen, Christian Theel, Dennis Uieß, Beate Weidenhammer, Christian Westermeier and Yilan Zhou for their direct contributions to the wikis, the creation of videos and exercises. And finally a thanks to all the students who have contributed with their works to the Teachwikis. Thanks also to Leslie Udvarhelyi for linguistic help.

REFERENCES (RÉFERENCES)

Härdle, W., Klinke, S., Müller, M. (2000) *XploRe Learning Guide*, Springer Verlag, Heidelberg, Germany.

Härdle, W., Klinke, S., Ziegenhagen, U. (2007) *On the Utility of E-Learning in Statistics*, International Statistical Review, 75, p. 355-364.

Klinke, S., Kuhlee, D., Theel, C., Wagner, C., Westermeier, C. (2009), *MM-Stat MultiMedia-Statistik: Statistische Datenanalyse webbasiert, interaktiv und multimedial*, SFB 649 Discussion paper SFB649DP2009-047, Humboldt-Universität zu Berlin, Berlin, Germany.

Klinke, S., Zlatkin-Troitschanskaia, O. (2007) *Embedding R in the Mediawiki*, SFB 649 Discussion Papers SFB649DP2007-061, Humboldt-Universität zu Berlin, Berlin, Germany.

Rönz, B., Müller, M., Ziegenhagen, U. (2000) *The Multimedia Project MM*STAT for Teaching Statistics*, COMPSTAT 2000 - Proceedings in Computational Statistics. Bethlehem and van der Heijden (eds.), Springer Verlag, Heidelberg, p. 409-414

Witzel, R., Klinke, S. (2002), *MD*Book online & e-stat: Generating e-stat Modules from Latex*, In: COMPSTAT 2002 - Proceedings in Computational Statistics - 15th Symposium held in Berlin (Germany) by W. Härdle and B. Rönz (eds.), Physika Verlag, Heidelberg, p. 449-454.

Wikipedia Kurier (2009), *Wikipedia und Statistik*, Ausgabe 09/2009

Weblinks

Electronic books	`http://fedc.wiwi.hu-berlin.de/xplore/ebooks/html/`
Project home in German Wikipedia	`http://mars.wiwi.hu-berlin.de/mediawiki/sk/index.php/Statistik_in_der_Wikipedia_-_Verbesserung_von_Qualität_und_Zugang`
LaTeX2html	`http://www.latex2html.org`
LatexWiki	`http://zwiki.org/LatexWiki`
Link cloud	`http://mars.wiwi.hu-berlin.de/mediawiki/wpstatde`
Mediawiki	`http://www.mediawiki.org`
MM-Stat I	`http://www.mhsg.de/index.php?id=110`
MM-Stat II	`http://www.mm-stat.org`
Statwiki	`http://statwiki.wiwi.hu-berlin.de`
Teachwiki	`http://teachwiki.wiwi.hu-berlin.de` and `http://mars.wiwi.hu-berlin.de/mediawiki/teachwiki2010`
Wikibook Statistics	`http://en.wikibooks.org/wiki/Statistics`

All web pages has been accessed between 9th and 30th April 2011.