

Online statistics labs

STEFANIE WULFF

Recent publications in the field of corpus linguistics (including several in this and the previous issue of CLLT) strongly indicate that the field is on its way from a view of corpora as mere repositories of authentic data from which examples can be culled *ad libitum* to a methodology that analyzes linguistic phenomena systematically and exhaustively as they manifest themselves in corpus data. Thereby, corpus linguistics is becoming an attractive complement to other empirical research methods in the language sciences, such as experimental designs.

One essential feature of any full-blown methodological approach is that its results are subjected to the scrutiny of statistical testing. However, while access to expensive statistics software packages is routinely available at most departments in the natural and social sciences, many linguistics departments are not as well equipped. In this review, I will briefly present and evaluate five freely accessible online statistics labs that will satisfy many of the statistical demands of a quantitatively-minded corpus linguist.

I use the term “online statistics labs” to refer to websites which offer interactive applets for the online calculation of different statistical procedures. That is, I specifically selected websites

- that do not require downloading some program to do the calculations or ask the user to enter data and then only return the results with some delay via email, but that allow the user to enter data *directly* into a web interface and that *immediately* supply the results;
- that deserve the name ‘laboratory’ in the sense that they are not restricted to one or two individual statistics, but let the user choose among the several *commonly used and established statistical procedures* (with a particular focus on the needs of corpus linguists) and ideally also allow further editing of the data, including various kinds of graphical display;

- that do not presuppose the user's detailed acquaintance with some programming language, but offer *user-friendly interfaces*;
- that offer their services *completely* and *for free*, that is, restricting neither the number of utilities made available nor the access to the output of the calculations depending on payments, be they direct or indirect payments (taking the form of, for instance, obligatory memberships or donations, participation in some kind of experiment, etc.).

In view of the significant number of websites offering calculations of statistics, it goes without saying that the final selection of statistics labs is non-exhaustive and subjective. However, the above-mentioned criteria for the selection of pages ensure that all sites presented here stand out in terms of user-friendliness and simplicity of use.

Rice Virtual Lab in Statistics

URL: <http://www.ruf.rice.edu/~lane/rvls.html>

This site was created and is maintained by David M. Lane, Associate Professor of Psychology, Statistics, and Management at Rice University, Texas. The site assembles a variety of very useful functions.

First, it is linked to Lane's *HyperStat Online Textbook* which comprises 18 chapters dealing with, among other issues, the analysis of univariate and bivariate data, sampling distributions, analyses of variance, or the chi-square test(s); each chapter is rounded off with exercises (and solutions!) to help users check their understanding.

Second, there is a link to Java applets which show interactive demonstrations of the statistical procedures and concepts, e. g. ANOVA, binomial distribution, effect size, regression, or t-test. Every demonstration is accompanied by exercises and instructions, in which the most relevant basic concepts (e. g., *normal distribution*) are linked to pages providing detailed explanations.

Third, by calling up *Case Studies*, the user can delve into authentic research examples from different disciplines. For every study, the website provides the raw data, background information on the research question, information on the experimental design, the materials employed, descriptive/inferential statistics, and the resulting interpretation.

Finally, and most importantly, the *Analysis Lab* link opens an interface which allows the user either to redo one of the afore-mentioned case studies or enter their own data for analysis. If one is not sure what the data should look like, the raw data samples of the corresponding case study will provide the answer. Once the data are entered and predictor,

dependent, and grouping variable(s) are specified, a variety of basic statistical procedures and expository functions are at the user's disposal, including descriptive statistics, box plots, *t*-tests and confidence intervals, histograms, one-way ANOVA, chi-square, and correlation/regression measures. The only drawback is that only Macintosh OSX users can copy-and-paste data into and out of the applets via a java AWT clipboard – everyone else has to enter them manually.

In sum, the *Rice Virtual Lab in Statistics* is not only the perfect starting point for statistical laymen who still need some orientation as to choosing the right procedure etc, but also a highly valuable source for teachers or other people who have a more advanced statistical knowledge and simply want to carry out a quick calculation.

Simple Interactive Statistical Analysis (SISA)

URL: <http://home.clara.net/sisa>

SISA was written by Daan Uitenbroek, research and statistical consultant in Hilversum, NL. *SISA* is an interesting collection of more than 130 modules that perform simple interactive statistical calculations via the internet. While *SISA* is not intended to be a complete statistical package, it includes a large number of functions for hypothesis testing and sample size calculations, many of which are difficult to find elsewhere. For instance, there is a form that computes the Fisher Yates exact test, which should be used instead of the more widely established chi-square test if 20% or more of the cells that enter into the calculation have a frequency of less than 5 – which is very often the case with linguistic phenomena. The individual web forms for entering data are very simple and clearly laid out; unfortunately, data cannot be imported, and the values for every cell have to be entered individually, which can be cumbersome if one wants to obtain results for more than one test. Also, there is no automatic saving function inbuilt into the applets, but the output for every procedure can be copied and pasted manually out of the form – so in this respect, it is a little bit easier to handle than, say, the output provided by the *Rice Virtual Lab in Statistics*.

There is a *Guide to Procedures* link for each procedure which provides information about the required input and mode of operation of this procedure. Where it is useful, Uitenbroek also summarizes ongoing discussions concerning the appropriateness of the individual procedures, and he provides relevant references on the topic and refers the user to possible alternatives.

Another feature of *SISA* is that it offers free downloads for some statistics that are particularly useful for linguistic analyses, a useful ser-

vice for users who do not have a permanent Internet access. There are Windows programs for multinomial and crosstabulation tests (for up to 2×7 or 3×3 tables) such as the above-mentioned Fisher Yates exact test, Odds Ratio, t -tests, and others; and several DOS programs for functions such as binomial, Poisson, or multinomial distributions, which, as the author admits, are only predecessors of the current *SISA* programs, but distinguish themselves in that they are much smaller and faster as compared to the HTML/Java-based scripts on the website.

To conclude, *SISA* is the right website for people who want to carry out some (difficult-to-get-elsewhere) calculations and cut-and-paste the output directly. Particularly for multiple tests, the free download programs offer a way out of (potentially) costly online calculations.

StatCrunch

URL: <http://www.statcrunch.com>

StatCrunch, developed under the guidance of R. Webster West at the Department of Statistics at the University of South Carolina, was explicitly created as an alternative to commercial statistics software packages and the problems these entail (in terms of charges, cryptic procedures only transparent for hard-core statisticians, etc.).

Perhaps the greatest advantage of *StatCrunch* as compared to other online statistics labs is that data (from text files or excel sheets) can be easily uploaded from and exported to a local computer or entered via copy-and-paste; moreover, they can be stored under the user's account, so results are available even if one does happen not to have access to one's personal computer.

The package includes a satisfactory assembly of the most widely used statistical procedures such as summary statistics, t -tests, regression, and nonparametric tests – although limitations become obvious once one looks for potential alternatives to these statistics. For example, the only analysis-of-variance procedure available is a one-way ANOVA, and the only available goodness-of-fit measure is the chi-square test. On the other hand, the user can choose among a respectable number of graphics to give their results the final expository polish.

Unfortunately, the *Help Topics* link cannot compare in quality to any of the afore-mentioned websites: it is restricted to a technical explanation of how to carry out the procedure in question in *StatCrunch*. Asking for information on *logistic regression*, a procedure not available by all online statistics labs, the user does not even get technical instructions, even though the term is highlighted in blue on the 'help topics' page.

In summary, despite its minor weaknesses, *StatCrunch* transcends other online labs in that it is a real all-in-one laboratory for data maintenance, analysis, and exposition.

VassarStats

URL: <http://faculty.vassar.edu/lowry/VassarStats.html>

The *VassarStats website* was created and is maintained by Richard Lowry, Professor of Psychology at Vassar College, NY. *VassarStats* offers a substantial list of different statistical procedures, including various kinds of probability measures, distribution generators, correlation and regression measures, analyses of variance, and many more.

For every procedure, detailed technical information on how to enter the data into the online applets as well as detailed mathematical explanations are provided. In addition, less technical help is available on the companion site *Concepts and Applications on Inferential Statistics*, an introductory statistics textbook featuring many interactive elements.

Generally, the data have to be entered manually as in *SISA*, but for some procedures (those that typically require larger data samples, e. g. linear regression), data can alternatively be entered via copy-and-paste. There is no automatic *save* function, and unlike in *SISA*, which displays all results in a single window so that one can copy the results en bloc, *VassarStats* requires the user to copy-and-paste results one by one from the interface.

There are a number of attractive extras which make it unnecessary to leave the website during a session, such as a notepad, a simple graph maker, and a pocket calculator. Additionally, two excel spreadsheets to calculate either a one-way ANCOVA or linear correlation and regression statistics can be downloaded free of charge, as well as several PDF documents (e. g., showing a table with the critical values of chi-square).

The *VassarStat* website is a very useful site for doing quick calculations, and as a combination of textbook and online lab, it is an ideal source for students who are looking for a single source to familiarize themselves with the theoretical concepts as well as the corresponding mathematical formula and simultaneously carry out calculations.

Hossein Arsham's website

URL: <http://home.ubalt.edu/ntsbarsh/Business-stat/home.html>

This site was created and is maintained by Prof. Hossein Arsham, Research Professor in Decision Science, Simulation, and Statistics at the

University of Baltimore, Maryland. Under *Instructional Activities*, the user can launch a variety of web-enhanced courses, many of which are also of potential interest for linguists.

First, the link *Excel for Statistical Data Analysis* allows the user to scroll their way through a step-by-step introduction beginning with the very basic steps (how to open a document or enter data into excel spreadsheets) and progressing to calculations of regressions, ANOVAs or chi-square tests.

Second, under *JavaScript E-labs Learning Objects*, there are around 40 online applets for descriptive statistics, probability computations, and monofactorial as well as multifactorial tests. Under *Requirements for most tests and estimations*, the user can take advantage of a number of applets which are difficult to find elsewhere – among other things, one can check whether a data set includes outliers, test one's data for normality or randomness, or determine optimal sample sizes. Every procedure is explained in clear and simple terms (since the applets are intended to accompany Arsham's courses, these explanations are, however, mostly very short indeed – but cf. below), and technical instructions are provided with respect to how the data have to be entered. The user cannot copy and paste the results, but has to note them down manually (or, of course, print them using their web-browser's print function).

Third, from the *Statistics Resources* link, the user can access a large collection of online sources relating to statistics, including different procedures, organizations, statistical software, and glossaries.

This website brings together a vast amount of different resources, and accordingly, its inner structure is quite complex in places. Moreover, for every major link, mirror sites are offered for Europe, Asia, and the Middle East, and more often than not, there is reference to a Spanish version of the site. Although this principally stands out as a rare and creditable service to non-English speaking users, it does affect the overall usability of the site somewhat. However, one of the topmost links provided on the main page is *A Searchable List of All My Web Sites*, by which the user can jump directly to a relevant page.

In sum, Arsham's website is recommendable for students and teachers alike. Particularly self-taught statisticians-to-be will value the multitude of available resources and the easily accessible instructions and explanations.

University of Bremen