# Guide To The Course - Introduction to Statistics

This course was created in view of its use in a fully on line class. However, its structure is also based on traditional face to face class that was taught a few quarters ago. Its content is licensed according to the CC-BY license:



Introduction to Statistics by Federico Marchetti - OLC project is licensed under a [Creative Commons Attribution 3.0 Unported License](http://creativecommons.org/licenses/by/3.0/).

Based on a work at [http://www.shoreline.edu/fmarchetti](http://www.shoreline.edu/fmarchetti/stats/index.html).

It relies extensively on parts taken from the Public Domain course, created by Rice University, together with the University of Houston, Clear Lake, and the University of Houston, Downtown, by David Lane (contact person), Joan Lu, Camille Peres, and Emiliy Zitek (with contributions by many others). While a version 2.0 is in progress (and available as beta), we will rely on the original version 1.0. It is available in its entirety at [http://onlinestatbook.com/index.html Online Statistics: An Interactive Multimedia Course of Study](http://onlinestatbook.com/index.html).

## Structure of the Course

In its present form, the course is articulated by modules, each assumed to take a week of class time. The schedule is the following:

1. Week 1. Familiarize the students with the structure of the course, and the tools that will be used
2. Week 2. Introduction to Descriptive Statistics
3. Week 3. Introduction to Probability 1: basic ideas
4. Week 4. Introduction to Probability 2: limit theorems
5. Week 5. Inferential Statistics 1: Interval Estimation
6. Week 6. Inferential Statistics 2: Tests - significance
7. Week 7. Inferential Statistics 3: Tests - power
8. Week 8. Inferential Statistics 4: Regression
9. Week 9. Applications: Examples and Sampling
10. Week 10. Review

Weeks 2 to 9 come with lecture material covering the topic of the week.

## Assessments

The suggested assessment tools include the following.

### Homework

Homework comes in two flavors, with the possibility of an additional project.

#### **Computational statistical homework**

Homework for weeks devoted to statistical topics (that is, excluding the two probability weeks), has students apply the specific tools introduced in that week (e.g., descriptive methods, parameter estimation, and so on) to a few data sets, provided by the course from the start. The idea is that the same data sets will be successively analyzed, each week with the tools of that week. These data are computer simulated data, so that they have, at least approximately, the theoretical features required by the theory for the application of statistical methods (e.g., two data sets are simulated normal, and one is simulated exponential). Additional, optional, data sets are available, simulated according to a Weibull and a Cauchy distribution. Applying statistical tools geared towards a normal distribution to data simulated under different assumptions should help in gaining a feeling for the scope and limitations of these tools.

The tool required to perform the calculations is a spreadsheet. This is motivated by the following arguments. In real life, this is how data is analyzed unless more powerful dedicated software is needed. It is also the only practical way to work with data sets that are not extremely small, and so not well suited for reliable statistical analysis. Another consideration is that the actual technicalities of the computations involved are extremely simple (they can all be performed on a simple non scientific calculator), and, in any case, are also addressed by the "pencil and paper" assignments (see below). Also, the learning curve required to perform these operations is a good investment into computer literacy in itself.

Hopefully, applying standard inferential methods to the normal data will yield results in good agreement with the parameters behind the simulation. Applying the same methods to data simulated under different assumptions should provide some experimental evidence of the reach of these methods.

#### **"Pencil and Paper" Assignments**

To complement the probability chapters, students are supposed to submit more traditional assignments about this part of the course. Additionally, as preparation to the exams (see below), a traditional assignment for Interval Estimation and Statistical Testing complements the computational assignment. While the calculation needed for these problems can be performed with the same spreadsheet used for the computational assignments, the data is presented in a form that allows the solutions to be found even with a very basic calculator and the standard statistical tables needed to evaluate distributions. This in view of the fact that any exam would be difficult to administer in a way that allowed the use of a spreadsheet, and also as a way of familiarizing students with ways to do statistics when only limited technology is available.

#### **Special Project**

Besides the data provided by the course, to be used for the computational assignment, students are strongly encouraged to find one or more real data sets about some topic of personal interest, and analyze these sets in the same way as the simulated sets. Whenever possible, doing this in small teams is the recommended way. This last suggestion was easily implemented in a face to face class, and was more challenging in an on line class, as was to be expected. The goal of this project is to reinforce the techniques we are trying to learn, with the additional motivation of learning about an issue the student is personally interested in.

### Tests

#### **Exams**

In the on line class, we also implemented two traditional exams (a Midterm and a Final), similar to the "pencil and paper" assignments. The data for the statistical part of the tests is presented with enough information so that no special calculators, beyond a very basic one, is needed, and the test sheets include the statistical tables needed to answer questions without access to a computer.

#### **Quizzes**

The course relied on the [WAMAP site](http://www.wamap.org) for a collection of on line quizzes to complement the activities listed above. In our implementation, the quizzes could be taken any number of times, and were meant as a participation activity, rather than a true assessment activity. The goal was to provide a means for simple self-testing, giving the option of repeating as needed, to acquire the necessary familiarity with the material.

## Outcomes

### Technical and Computational Skills

These skills pertain to the ability to apply formulas and methods to specific statistical goals. These skills are of relatively limited importance for a number of reasons. While it is always better to have a clear concept of what is involved in calculations you are performing, it is a fact that most of these are now pre-programmed in software, requiring much less manual ability from the operator than in the past.

Hand (or even calculator) based calculations are only feasible for really small data sets, where, incidentally, the meaningfulness of these calculations is often open to debate. In fact, in those circumstances, it is definitely appropriate to apply first a critical analysis of the problem, and only afterwards proceed to apply standard methods, if appropriate.

Large data sets, where it is easier to argue for the validity of standard methods, mostly based on limit laws, can only be reasonably handled through computer programs. "Real" statistical packages, both commercial (S, SPSS, ...), and open source (such as R), essential when dealing with very large data sets, are beyond the scope of an introductory course. On the other hand, fairly large data sets are easily handled by spreadsheets which are, by and large, the standard tool for data sets of this size. Since practically all standard analysis tools are hard-coded in most spreadsheets (and the open source spreadsheet Gnumeric is especially well endowed in this area), the actual technical skill required to operate meaningfully is relatively limited.

For this reason, while the ability to operate the software on specific data is definitely an expected outcome of the course, it is not its primary focus.

### Critical evaluation skills

This is what should be the goal in this class. Most methods, both "descriptive" and "inferential", assume, openly or implicitly, a model behind the data. The appropriateness of the model is of critical importance, and should definitely be brought in the open. To be clear: data don't speak by themselves. For example, a numerical data set is only a list of numbers. Any conclusion we may draw is based on additional assumptions that we are making, and it is best if these assumptions are made explicit. Even problems, like polls, that are by now well established, rely on assumptions that are often very delicate to verify. We will strive to discuss as extensively and clearly as possible, what these assumptions imply, and how a given experiment could fail to meet them.

A connected issue is the understanding of what common statistical conclusions really imply."Statistical Tests" are a prime example of not so obvious meaning of their outcomes, but similar comments can be made about any estimation technique.

There are limits to how far this issue can be pursued in an introductory course. It can easily happen that students might concentrate exclusively on the technicalities and overlook th discussion about the tools. Also, the critical examination of the limits of validity of the tools is inevitably somewhat generic (quantitative evaluations are research material). One area of development of this course would be the creation of additional specific assessment tools to help a better understanding of this issues.

### Assessments and Outcomes

The assignments based on spreadsheet analysis should allow students to acquire basic data manipulation skills. Including work on non normal data (the basic data file includes a simulation form an exponential distribution) is meant to provide a chance to compare the output form "standard" tools, that assume normality. Of course, it is possible to analyze exponential data rigorously, without recourse to a normal approximation, and, if time permits, that would provide a valuable example of how approximate and precise tools will perform.

The "pencil and paper" assignments, as well as the two exams, are meant to familiarize students with traditional table-based estimations. While mostly unnecessary, they are somemes convenient, or sitimes, in special circumstances, convenient, or simply the only available tools. Moreover, they complement the "automatic" operation that sometimes spreadsheets offer, by forcing more traditional "manual" manipulations, that, again, are sometimes necessary.

The critical skills that are an important goal are better developed in an interactive context, with at least some live interaction between instructor and students. This allows to put appropriate emphasis on the conceptual underpinnings of the methods, as they are introduced. In a more detached situation, as in a fully on line setting, this is harder to effect, as texts can push this issue only so far. The presence of data that is not normal by design (besides the "advanced" data sets, the exponential simulation in the basic data set, and the non normal simulation in the regression data set) is a tool to encourage students to confront the results of normal-based tools when the situation does not reflect the assumptions. In the course of the class, this was strikingly successful when the non normal regression data was analyzed by the students: the high correlation coefficient, coupled with an excessive sum of residuals was a surprise to most, which, with appropriate follow up by the instructor, has provided a concrete example of the limits of our tools.

## Format Of The Material

All the material in the course is available in digital, open formats. Specifically,

* Plain text content is presented in plain HTML
* The On Line Stat Book content is also in plain HTML, but is also available in PDF format (the course includes a compressed file with all the 12 chapters we refer to)
* Text including mathematical formulas is presented both in XHTML (HTML5 compliant, using MathML for mathematical rendering), and PDF. This double format is motivated to allow for files with possible links (XHTML), but that may not be well handled by some browsers (notably Internet Explorer), so that files without the live links, but easy to access (PDF), may be used as an alternative
* Spreadsheet files are available in native Gnumeric format (see the tools section for motivation), as well as Open Document Format (\*.ods files), and Excel (\*.xls) format, which is not open but widely used.

## Tools Required For Students

The main tools to access this course are

* A browser. Firefox (an open source browser) and Opera (closed source, but free) should be able to handle the MathML content of the course. Other browsers may have varying levels of trouble with MathML
* A PDF reader. The most common reader is Adobe Acrobat Reader (free, but closed source), but there is an array of open source PDF readers that work very well across multiple platforms (for example, [Okular](http://okular.kde.org))
* A spreadsheet. Any modern spreadsheet has all the necessary statistical functions already coded. The most common open source spreadsheet is probably the Open Office/Libre Office spreadsheet. An important alternative is the Gnumeric spreadsheet which has additional features, not available in any other similar program, that make statistical analysis particularly easy. Assignments relying on spreadsheet analysis would be best turned in in their native form.
* "Pencil and Paper" assignments are supposed to be turned in traditional ways. In a face to face class, paper submissions are the obvious choice. In an on line class, care should be taken to avoid submission in special formats that may not be readable by the intended audience. Thus, word processor files are discouraged, especially if they include special fonts and/or rely on formula editors that may not be available on the other side. PDF or text files seem the best choice for computer generated submissions, while scans or pictures of handwritten work are also a convenient form.

## Tools Used For Creating This Course

This course was prepared on different computers, all running a version of Linux, and using exclusively open source (mostly GPL-licensed) software. All packages used are available in the main Linux repositories, but have versions running on Windows as well, sometimes through the services of Cygwin, a Unix-like environment for Windows. I am not familiar with the availability of open source software under Mac OS X, and cannot provide information on this.

### Pure Text Editing

Pure text is provided in HTML format. To minimize the hand coding we relied mainly on [Amaya](http://www.w3.org/Amaya), a browser/HTML open source WYSIWYG editor from the W3C consortium and the French INRIA. Advantages are its strict enforcement of web standards, and its resulting very clean HTML pages. Of course, there are innumerable other HTML editors, especially for those familiar with the language. A basic decision was to stay away from word processors, as they tend to create unnecessarily bloated pages.

### Text and Mathematical Formulas

Amaya supports MathML (an HTML5 standard), but I have had mixed success in getting other browsers recognize it. For this purpose I relied on [Texmacs](http://www.texmaxs.org), a remarkable math editor, available on Linux and, thanks to [the Cygwin package](http://www.cygwin.com), on Windows. The editor uses its own XML format (\*.tm files, that are included in the instructor distribution), but produces frugal XHTML, with MathML elements that are well understood by updated browsers (Firefox 6.x, Opera 11.x read it natively, IE needs a plugin). It also exports well to LaTeX, and produces good PDF files. A bug in its current version causes the "export to PDF" command to produce PostScript, instead of PDF (PDF is, basically, compressed PostScript). While inconvenient, this bug is not too disruptive: running the output through a program like ps2pdf generates perfectly good PDF files. It also exports well to LaTeX (LaTeX exports are also included in the instructor distribution)

[Lyx](http://www.lyx.org) is another math editor with great LaTeX and PDF output. It also handles pictures in a slightly easier way than Texmacs, and is possibly overall easier to use. As of its current version, its mathematical HTML output still relies on external image files, and is hence less convenient than Texmacs, in this respect. Some files presented only in PDF and LaTeX format were created in Lyx.

I have stayed away from generic word processors, as their mathematical output is always of inferior quality, except for some assignments and tests, where formulas were not needed. The two word processors used were LibreOffice 3.3 Writer and Abiword (both available on most platforms). The motivation was the easy way they could accommodate tables generated by their spreadsheet cousins (respectively, LibreOffice 3.3. Calc and Gnumeric), part of the same office suites.

### Graphics

Graphics is mostly handled by the spreadsheet this course is based on. For additional pictures, function graphing has relied on [Kmplot](http://www.kde.org/kmplot.html) a very good scientific graphing package, and, occasionally, on [Lybniz](http://www.sourceforge.net/lybniz), a more Spartan, but convenient graphing package. For raster (bitmap) manipulations, as much as it may be overkill, it's hard to beat [The Gimp](http://www.gimp.org).

### Slide Shows

Time constraints have prevented, for now, a wider creation of slide shows to complement the lecture material. The only one included at this stage was created using Open Office/Libre Office Impress, which is certainly a very capable presentation creator. This is presented as an HTML slide sequence, so no slide software is needed to read it.

There is another slide show, a very short introduction to probability, which was not deployed in the course, as is included as an Impress file. For future additions, I was happy to learn of other, lighter, open source tools, which may lack some of the features of more complete presentation software, but have the notable advantage of being able to turn a sequence of PDF files into a slide show.

### Mathematical Software

The course is based on using a spreadsheet for all statistical calculations. This with the idea that that is the most common situation encountered in working environments without need for heavy duty analysis. If a more professional tool was required, the open source statistical suite R would be the obvious choice, and the IPSUR project has a well thought out course, based entirely on R. R implements the S language, popularized by the commercial package S-Plus, but there is also an open source project implementing the SPSS language, if that is preferred.

Any spreadsheet will work for this course, but the one we are referring to specifically is [Gnumeric](http://www.gnumeric.org). This is a relatively "light-weight" open source spreadsheet (it has fewer bells and whistles than other better known packages), but comes with many more hard-coded statistical functions than any competitor. While this is mostly a convenience for students, the need for generating simulated data sets with different distributions made this a must use for creating the course. Gnumeric comes with pre-programmed random number generators for many different distributions, while all other spreadsheets only come with the basic, uniform random number generator. It is obviously possible to generate any other distribution form this, but this is a delicate and time consuming chore - not all methods are created equal, and the best ones are fairly complex - so that having them already available makes a big difference (for example, the normal distribution is generated by Gnumeric using one of the most reliable current methods, which is more complicated than the simple algorithms usually presented in entry level probability courses).