# USE OF EXCEL IN A FIRST COURSE IN STATISTICS FOR MATHEMATICS STUDENTS

Lucette Carter and Mathilde Mougeot, University of Paris X – Nanterre, France

*The paper describes the use of EXCEL in the teaching of descriptive statistics to second year students working for a mathematical degree, oriented towards Human Sciences (Economics or Psychology) at the University of Paris X-Nanterre.*
*EXCEL capabilities and its limited number of integrated functions - which encourages the students to explore for themselves the intermediate steps of the various statistical methods - have helped the students to attain a deeper understanding of the statistical tools, particularly in the case of more advanced topics like analysis of contingency tables or multiple regression.*

## INTRODUCTION

This paper is based on experience with a course of descriptive statistics for second year students working for a mathematical degree, oriented towards Human Sciences (Economics or Psychology). In its original version, this teaching course was backed up by a statistical package. (RATS was chosen in this particular instance for its focus on econometrics and time series.) The advantages of this included enhancement of the students' motivation and improvement of their understanding and participation. However these advantages were counterbalanced by two main disadvantages. The first was that too much time was needed for getting acquainted with the necessary technicalities (the contents of the package menus, the way to handle data files, etc). The second disadvantage was a tendency to consider the computer as a "blackbox" between the data input and the results output.

To avoid these inconveniences, we developed a new version of the course in which the statistical package (RATS) was replaced by a spreadsheet. We chose EXCEL in this particular instance since it is now available on any personal P.C. This modification was suggested by the successful results of previous experience with integration of a spreadsheet at an introductory level in statistic (Hunt (1994)).

Although most students work in pairs during the computer sessions, the final assessment is based on an examination involving individual "real time" work on a given, previously unseen, problem.

## TEACHING/LEARNING ENVIRONMENT

This first course in statistics covered the following topics: univariate and bivariate statistical distributions, multidimensional contingency tables (with emphasis on two dimensions), simple and multiple regression, time series, and introduction to principal components analysis. The program relied on the mathematical background of the students (who already had some acquaintance with multidimensionnal Euclidean space and matrix algebra). It was designed to meet the interests of students oriented towards human sciences such as psychology and sociology (for whom the structure of contingency tables are essential) and also students oriented towards economics (for whom regression and time series are more important). The course consisted of 24 sessions of 2 hours each, in which the presentation of statistical methods was followed by worked applications ("Travaux Dirigés"). Class room exercises were carried out using just pocket calculators, so they were restricted to a small amount of data. The use of EXCEL as a statistics assistant was introduced at the outset, in three tutorial classes of one hour each. These tutorial classes (in groups of 15 students, with one P.C. each) were enough to "launch" the students, so that they could then work independently (using collective or private computing facilities) on more advanced problems of numerical and graphical analysis during the later stages of the course.

## FIRST CONTACT WITH EXCEL AS A STATISTICS TUTOR

When they had mastered the basic elements (such as absolute and relative cell addresses and the capability of "copying" formulae) of EXCEL, the students were asked to calculate parameters (such as average, variance, standard deviation, covariance, correlation coefficient, medians, ...) for raw data. One attraction of EXCEL for the student is the simplicity of data handling and acquisition. EXCEL facilitates progressive learning of statistical tools, which are first written explicitly (using only the basic arithmetic operators and the command such as copying a formula built in one cell to another cell in the same row or column) before the introduction of the related integrated commands (such as VARIANCE, COVARIANCE,...). The immediate possibility of checking all calculations by a backward process makes it easy to trace any error back to its origin.

When the students are made familiar with the statistical formulae, they are presented with the corresponding EXCEL integrated functions such as AVERAGE, VARIANCE,...

The visibility of the data, of the calculation steps, and of the final output provides the opportunity of checking in “real-time” the impact of data modification or changing of a variable scale unit on the distribution parameters

## ANALYSIS OF GROUPED DATA AND CONTINGENCY TABLES

The analysis of grouped data distributions was emphasised in the early stages of the course, and some time and effort was devoted to the study of contingency tables of 2 or higher dimensions. These contingency tables appear in the study of large data sets resulting from opinion polls, clinical essay, demographic surveys, and are of particular interest for future working fields of many of our students.

EXCEL has no specific integrated function that can directly produce numerical and graphical summaries of grouped data. This limitation is in fact a key pedagogical advantage, since the students must write the relevant intermediate formulae explicitly. We illustrate this mechanism by some examples in one and two dimensions.

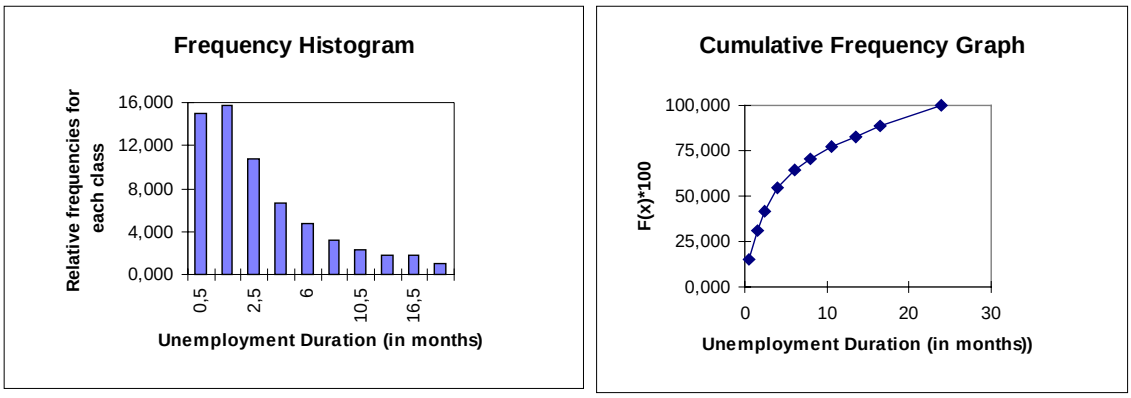



*Figure 1*

Figure 1 displays the distribution $\{(e_{j-1}, e_j , n_j), j=1...10\}$ of unemployment duration (in months) of a sample of n=14 845 inscribed at the ANPE (Agence nationale pour l’emploi) in 1986 (Cardellini, 1993). The average duration $\bar{x}$ ($\bar{x}$ =7,33 months) is calculated in 3 steps: contribution of the first class: $f_j*x_j$, contribution of the subsequent classes (by the copying process); sum of all contributions $\Sigma_j f_j*x_j$. A similar process is applied for the calculation of the variance. At this stage, the students are presented with the EXCEL mathematical function SUMPROD corresponding to the Cartesian scalar product which is very useful for the calculation of uni and bivariate distributions parameters: it provides a good pedagogical intermediate between longer explicit calculations and the use of integrated function. For example Mean(X) = SUMPROD ($n_j$

column; $x_j$ column). The example in Table 1 presents the joint distribution $\{(x_j,y_k,n_{jk}), j=1,...,9;k=1,...,7\}$ of the age of the family head (variable X) and of the number of children under 16 years (variable Y).

Table 1

| | | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 22,5 | 27,5 | 32,5 | 37,5 | 42,5 | 47,5 | 52,5 | 60 | 70 |
| y1 | 0 | 1500 | 2700 | 2000 | 1900 | 2300 | 3000 | 6300 | 17000 | 22000 |
| y2 | 1 | 400 | 2100 | 2400 | 2100 | 2200 | 2500 | 2400 | 2500 | 1000 |
| y3 | 2 | 100 | 1200 | 2500 | 2600 | 2100 | 1300 | 1100 | 900 | 400 |
| y4 | 3 | 30 | 400 | 1300 | 1700 | 1300 | 700 | 400 | 300 | 200 |
| y5 | 4 | 0 | 100 | 600 | 900 | 700 | 300 | 200 | 100 | 0 |
| y6 | 5 | 0 | 40 | 200 | 400 | 300 | 100 | 50 | 100 | 30 |
| y7 | 6 | 0 | 0 | 200 | 300 | 300 | 100 | 50 | 0 | 100 |

The calculation of the covariance $\Sigma_{\{j=1..9\}} \Sigma_{\{k=1..7\}} f_{jk} x_j y_k - \bar{x}\ \bar{y}$ with a minimum of commands reinforces the student's technical ability. The quantity $\Sigma_{\{j=1..9\}} \Sigma_{\{k=1..7\}} f_{jk} x_j y_k - \bar{x}\ \bar{y}$ is calculated in 3 steps: contribution of the table first column $\Sigma_{\{k=1..7\}} n_{1k} x_1 y_k$ (using the SUMPROD command with appropriate relative and absolute addresses); contribution of the subsequent columns; sum of contributions. The calculation of covariance and of the correlation coefficient (equal to -0,43 in this case) is then straightforward. To assess the relation between the 2 variables, it is useful to calculate the conditional distributions for each variable and the conditional averages of Y given $X=x_j$ which are displayed in fig. 2

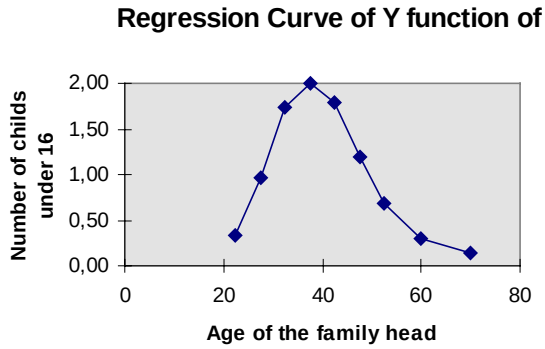



*Figure 2*

SIMPLE AND MULTIPLE REGRESSION

The concept of data modelling was introduced through the analysis of the scatterplots of various sets of observed data *{($x_i$,$y_i$), i=1,...,n}*. After the simple case of the

linear model $Y=a + bX$ where the coefficients are estimated by the least squares method (linear regression of Y on X), the students were presented with various intrinsically linear models $Y=f(X)$ (which can be transformed in a linear model through adequate transformations of variables). Generalisation to the case of several explanatory variables required a good knowledge of vector, matrix calculus and Euclidean n-dimensional geometry.

The steps to estimate the coefficients of the linear model $y=a0 + \Sigma_{\{j=1,...,p\}} a_j X_j$ were the following: construction of the explanatory variables variance/covariance matrix $V_X$ (which is then inversed using the EXCEL command INVERSE.MATRIX) and the column matrix $V_{YX}$ whose $j^{th}$ element is equal to $cov(Y,X_j)$; calculation of the inverse $V_X^{-1}$ of the matrix $V_X$ (using the EXCEL command INVERSE.MATRIX); calculation of the matrix product $V_X^{-1} V_{XY}$ to obtain the resulting coefficients column matrix A. The regression determination coefficient was calculated using the expression: $R^2 = A^t V_{XY} / Var(Y)$.

Once the students were made familiar with these techniques, they were introduced to the EXCEL command LINE.REG which directly provides the coefficients for simple or multiple regression. But the direct output of EXCEL for regression is restricted to the command LINE.REG and the command producing the coefficient of multiple determination $R^2$. Therefore, for each exercise, the students were asked to calculate estimated values of the variable Y and the corresponding normalized residuals. They could then produce the graphical displays to assess the fit of the model and detect possible outliers.

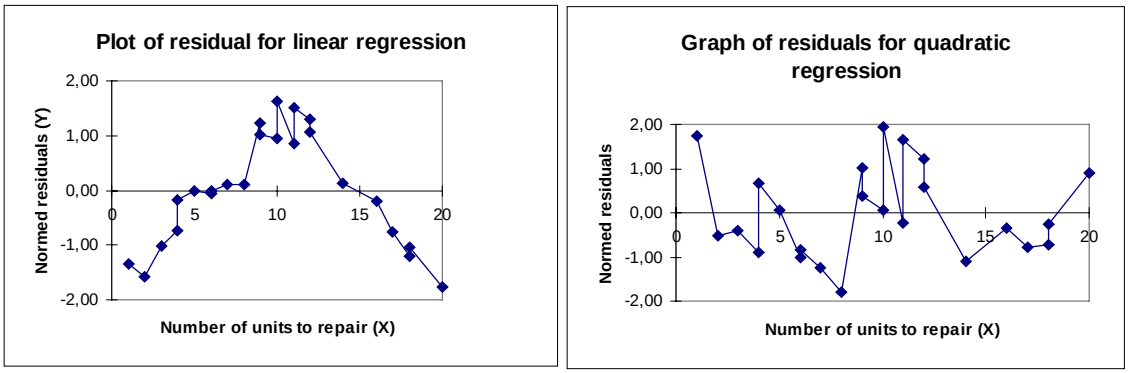
The example (Chatterjee and Price, 1991) in Table 2 used data for n=24 computer repairs, each characterised by a pair of variables (X,Y), where X represents the number of individual parts to be repaired and Y the duration (in minutes) of the job.

Table 2

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Units X | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 6 | 7 | 8 | 9 | 9 | 10 |
| DurationY | 23 | 29 | 49 | 64 | 74 | 87 | 96 | 97 | 109 | 119 | 149 | 145 | 154 |
| Units X | 10 | 11 | 11 | 12 | 12 | 14 | 16 | 17 | 18 | 18 | 20 | | |
| DurationY | 166 | 162 | 174 | 180 | 176 | 179 | 193 | 193 | 195 | 198 | 205 | | |

The results of a linear regression of Y on X and the graph of the residuals revealed the good fit of the model for the first 14 data. When the complete set of data is taken into account, the residual plot (Figure 3) exhibits a “non random” distribution which confirms that the linear model is no longer adequate. Figure 4 shows the residual plot

corresponding to the quadratic model $Y=a + bX + cX^2$, which confirms the goodness of fit of this model.




*Figures 3 and 4*

The EXCEL command LINE.REG can be used only for raw data. For bivariate grouped data (often presented in the form of a contingency table) the students had to calculate the coefficients of the chosen model explicitly. As we have seen, in the example of the contingency table in the previous section, a preliminary analysis of the table structure is necessary for guessing which model might be most appropriate.

## CONCLUSION

The use of EXCEL helped the students to attain a deeper understanding of the statistical tools and a greater ability to process the various basic methods, in particular in the analysis of contingency tables and linear (simple and multiple) regression. The key pedagogical interest of EXCEL is that it has limited integrated statistical functions so that the students must themselves elaborate the steps to arrive at the final results. However this process is quick and transparent, with the possibility of going backward to check the results (final or intermediate) at any stage of the statistical process.

Although the approach of this introductory course was technique-driven (rather than problem-driven) the capabilities of EXCEL (notably its “real-time” impact of data modification on the output, both numerical and graphical displays, minimal prerequisite to produce graphs,...) helped the students to get a better “feel” of the data and of certain statistical analyses. Hasty interpretation of results, and the too ready acceptance of any output, is sometimes induced by an unprepared use of traditional “black box” type statistical packages. One hopes that the data treatment with EXCEL will have avoided this danger by inducing a certain caution in interpreting results and an attention to prerequisites such as a sufficient amount of data.

REFERENCES

Cardellini, E. (1993). *Modèles Economiques de durée de chômage.* Mémoire de D.E.A., Paris X-Nanterre University.

Chatterjee, S. and Price, B. (1991). *Regression Analysis by Examples* .John Wiley and Sons.

Hunt, N. (1994). Teaching Statistics Using a Spreadsheet. In :ISI Publications (Ed. National Organizing Committee of ICOTS IV*) Proceedings of the Fourth International Conference on Teaching Statistics 2*.

Smyth, G. (1991). Computers and Computing in Statistics Courses. In :ISI Publications D. Vere-Jones (Ed). *Proceedings of the Third International Conference on Teaching Statistics, 2,* 144.