# 外文翻译稿和原稿

## 外文资料翻译—原文部分

**Using Excel for Statistical Data Analysis - Caveats**

**At A Glance**

We used Excel to do some basic data analysis tasks to see whether it is a reasonable alternative to using a statistical package for the same tasks.  We concluded that Excel is a poor choice for statistical analysis beyond textbook examples, the simplest descriptive statistics, or for more than a very few columns.  The problems we encountered that led to this conclusion are in four general areas:

* Missing values are handled inconsistently, and sometimes incorrectly.
* Data organization differs per analysis, forcing you to reorganize your data in many ways if you want to do many different analyses.
* Many analyses can only be done on one column at a time, making it inconvenient to do the same analysis on many columns.
* Output is poorly organized, sometimes inadequately labeled, and there is no record of how an analysis was accomplished.

Excel is convenient for data entry, and for quickly manipulating rows and columns prior to statistical analysis. However, when you are ready to do the statistical analysis, we recommend the use of a statistical package such as SAS, SPSS, Stata, Siesta or Minitab.

**Introduction**

Excel is probably the most commonly used spreadsheet for PCs. Newly purchased computers often arrive with Excel already loaded. It is easily used to do a variety of calculations, includes a collection of statistical functions, and a Data Analysis Tool Pak. Thus, if you suddenly find you need to do some statistical analysis, you may turn to it as the obvious choice. We decided to do some testing to see how good Excel would serve as a Data Analysis application.

To present the results, we will use a small example. The data for this example is fictitious. It was chosen to have two categorical and two continuous variables, so that we could test a variety of basic statistical techniques. Since almost all real data sets have at least a few missing data points, and since the ability to deal with missing data correctly is one of the features that we take for granted in a statistical analysis package, we introduced two empty cells in the data:

Each row of the spreadsheet represents a subject. The first subject received Treatment 1, and had Outcome 1. X and Y are the values of two measurements on each subject. We were unable to get a measurement for Y on the second subject, or on X for the last subject, so these cells are blank. The subjects are entered in the order that the data became available, so the data is not ordered in any way.

We used this data to do some simple analyses and compared the results with a standard statistical package. The comparison considered the accuracy of the results as well as the ease with which the interface could be used for bigger data sets - i.e. more columns. We used SPSS as the standard, though any of the statistical packages OIT supports would do equally well for this purpose. In this article when we say "a statistical package," we mean SPSS, SAS, STATA, SYSTAT, or Minitab.

Most of Excels statistical procedures are part of the Data Analysis tool pack, which is in the Tools menu. It includes a variety of choices including simple descriptive statistics, t-tests, correlations, 1 or 2-way analysis of variance, regression, etc. If you do not have a Data Analysis item on the Tools menu, you need to install the Data Analysis Tool Pak.  Search in Help for "Data Analysis Tools" for instructions on loading the Tool Pak.

Two other Excel features are useful for certain analyses, but the Data Analysis tool pack is the only one that provides reasonably complete tests of statistical significance. Pivot Table in the Data menu can be used to generate summary tables of means, standard deviations, counts, etc. Also, you could use functions to generate some statistical measures, such as a correlation coefficient. Functions generate a single number, so using functions you will likely have to combine bits and pieces to get what you want. Even so, you may not be able to generate all the parts you need for a complete analysis.

Unless otherwise stated, all statistical tests using Excel were done with the Data Analysis Tool Pak. To check a variety of statistical tests, we chose the following tasks:

* Get means and standard deviations of X and Y for the entire group, and for each treatment group.
* Get the correlation between X and Y.
* Do a two-sample t-test to test whether the two treatment groups differ on X and Y.
* Do a paired t-test to test whether X and Y are statistically different from each other?
* Compare the number of subjects with each outcome by treatment group, using a chi-squared test.

These tasks are routine for a data set of this nature, and all of them could be easily done using any of the above listed statistical packages.

**General Issues**

**Enable the Analysis Tool Pak**

The Data Analysis Tool Pak is not installed with the standard Excel setup.  Look in the Tools menu.  If you do not have a Data Analysis item, you will need to install the Data Analysis tools.  Search Help for "Data Analysis Tools" for instructions.

**Missing Values**

A blank cell is the only way for Excel to deal with missing data.  If you have any other missing value codes, you will need to change them to blanks.

**Data Arrangement**

Different analyses require the data to be arranged in various ways.  If you plan on a variety of different tests, there may not be a single arrangement that will work.  You will probably need to rearrange the data several ways to get everything you need.

**Dialog Boxes**

Choose Tools/Data Analysis, and select the kind of analysis you want to do.  The typical dialog box will have the following items:  
            Input Range:  Type the upper left and lower right corner cells. e.g. A1:B100.  You can only choose adjacent rows and columns.  Unless there is a checkbox for grouping data by rows or columns (and there usually is not), all the data is considered as one glop.  
            Labels - There is sometimes a box you can check off to indicate that the first row of your sheet contains labels.  If you have labels in the first row, check this box, and your output MAY be labeled with your label.  Then again, it may not.   
            Output location - New Sheet is the default.  Or, type in the cell address of the upper left corner of where you want to place the output in the current sheet.  New Worksheet is another option, which I have not tried.  Ramifications of this choice are discussed below.  
            Other items, depending on the analysis.

**Output location**

The output from each analysis can go to a new sheet within your current Excel file (this is the default), or you can place it within the current sheet by specifying the upper left corner cell where you want it placed.  Either way is a bit of a nuisance.  If each output is in a new sheet, you end up with lots of sheets, each with a small bit of output.  If you place them in the current sheet, you need to place them appropriately; leave room for adding comments and labels; changes you need to make to format one output properly may affect another output adversely.  Example:  Output from Descriptive has a column of labels such as Standard Deviation, Standard Error, etc.  You will want to make this column wide to be able to read the labels.  But if a simple Frequency output is right underneath, then the column displaying the values being counted, which may just contain small integers, will also be wide.

**Results of Analyses**

**Descriptive Statistics**

The quickest way to get means and standard deviations for an entire group is using Descriptive in the Data Analysis tools. You can choose several adjacent columns for the Input Range (in this case the X and Y columns), and each column is analyzed separately. The labels in the first row are used to label the output, and the empty cells are ignored. If you have more, non-adjacent columns you need to analyze, you must repeat the process for each group of contiguous columns. The procedure is straightforward, can manage many columns reasonably efficiently, and empty cells are treated properly.

To get the means and standard deviations of X and Y for each treatment group requires the use of Pivot Tables (unless you want to rearrange the data sheet to separate the two groups).  After selecting the (contiguous) data range, in the Pivot Table Wizard's Layout option, drag Treatment to the Row variable area, and X to the Data area.  Double click on “Count of X” in the Data area, and change it to Average.  Drag X into the Data box again, and this time change Count to Steve.  Finally, drag X in one more time, leaving it as Count of X.  This will give us the Average, standard deviation and number of observations in each treatment group for X.  Do the same for Y, so we will get the average, standard deviation and number of observations for Y also.  This will put a total of six items in the Data box (three for X and three for Y). As you can see, if you want to get a variety of descriptive statistics for several variables, the process will get tedious.

A statistical package lets you choose as many variables as you wish for descriptive statistics, whether they are contiguous. You can get the descriptive statistics for all the subjects together, or broken down by a categorical variable such as treatment. You can select the statistics you want to see once, and it will apply to all variables chosen.

**Correlations**

Using the Data Analysis tools, the dialog for correlations is much like the one for descriptive - you can choose several contiguous columns, and get an output matrix of all pairs of correlations. Empty cells are ignored appropriately. The output does NOT include the number of pairs of data points used to compute each correlation (which can vary, depending on where you have missing data), and does not indicate whether any of the correlations are statistically significant. If you want correlations on non-contiguous columns, you would either must include the intervening columns, or copy the desired columns to a contiguous location.

A statistical package would permit you to choose non-contiguous columns for your correlations. The output would tell you how many pairs of data points were used to compute each correlation, and which correlations are statistically significant.

**Two-Sample T-test**

This test can be used to check whether the two treatment groups differ on the values of either X or Y. To do the test, you need to enter a cell range for each group. Since the data were not entered by treatment group, we first need to sort the rows by treatment. **Be sure to take all the other columns along with treatment, so that the data for each subject remains intact**. After the data is sorted, you can enter the range of cells containing the X measurements for each treatment. Do not include the row with the labels, because the second group does not have a label row. Therefore, your output will not be labeled to indicate that this output is for X. If you want the output labeled, you must copy the cells corresponding to the second group to a separate column, and enter a row with a label for the second group. If you also want to do the t-test for the Y measurements, you’ll need to repeat the process. The empty cells are ignored, and other than the problems with labeling the output, the results are correct.

A statistical package would do this task without any need to sort the data or copy it to another column, and the output would always be properly labeled to the extent that you provide labels for your variables and treatment groups. It would also allow you to choose more than one variable at a time for the t-test (e.g. X and Y).

**Paired t-test**

The paired t-test is a method for testing whether the difference between two measurements on the same subject is significantly different from 0. In this example, we wish to test the difference between X and Y measured on the same subject. The important feature of this test is that it compares the measurements within each subject. If you scan the X and Y columns separately, they do not look obviously different. But if you look at each X-Y pair, you will notice that in every case, X is greater than Y. The paired t-test should be sensitive to this difference. In the two cases where either X or Y is missing, it is not possible to compare the two measures on a subject. Hence, only 8 rows are usable for the paired t-test.

When you run the paired t-test on this data, you get a t-statistic of 0.09, with a 2-tail probability of 0.93. The test does not find any significant difference between X and Y. Looking at the output more carefully, we notice that it says there are 9 observations. As noted above, there should only be 8. It appears that Excel has failed to exclude the observations that did not have both X and Y measurements. To get the correct results copy X and Y to two new columns and remove the data in the cells that have no value for the other measure. Now re-run the paired t-test. This time the t-statistic is 6.14817 with a 2-tail probability of 0.000468. The conclusion is completely different!

Of course, this is an extreme example. But the point is that Excel does not calculate the paired t-test correctly when some observations have one of the measurements but not the other. Although it is possible to get the correct result, you would have no reason to suspect the results you get unless you are sufficiently alert to notice that the number of observations is wrong. There is nothing in online help that would warn you about this issue.

Interestingly, there is also a TTEST function, which gives the correct results for this example. Apparently, the functions and the Data Analysis tools are not consistent in how they deal with missing cells. Nevertheless, I cannot recommend the use of functions in preference to the Data Analysis tools, because the result of using a function is a single number - in this case, the 2-tail probability of the t-statistic. The function does not give you the t-statistic itself, the degrees of freedom, or any number of other items that you would want to see if you were doing a statistical test.

A statistical package will correctly exclude the cases with one of the measurements missing, and will provide all the supporting statistics you need to interpret the output.

**Crosstabulation and Chi-Squared Test of Independence**

Our final task is to count the two outcomes in each treatment group, and use a chi-square test of independence to test for a relationship between treatment and outcome.  To count the outcomes by treatment group, you need to use Pivot Tables. In the Pivot Table Wizard's Layout option, drag Treatment to Row, Outcome to Column and to Data.  The Data area should say "Count of Outcome" – if not, double-click on it and select "Count". If you want percent’s, double-click "Count of Outcome", and click Options; in the “Show Data As” box which appears, select "% of row".  If you want both counts and percent’s, you can drag the same variable into the Data area twice, and use it once for counts and once for percent’s.

Getting the chi-square test is not so simple, however. It is only available as a function, and the input needed for the function is the observed counts in each combination of treatment and outcome (which you have in your pivot table), and the expected counts in each combination. Expected counts? What are they? How do you get them? If you have sufficient statistical background to know how to calculate the expected counts, and can do Excel calculations using relative and absolute cell addresses, you should be able to navigate through this. If not, you’re out of luck.

If you surmounted the problem of expected counts, you can use the Chicest function to get the probability of observing a chi-square value bigger than the one for this table. Again, since we are using functions, you do not get many other necessary pieces of the calculation, notably the value of the chi-square statistic or its degrees of freedom.

No statistical package would require you to provide the expected values before computing a chi-square test of independence. Further, the results would always include the chi-square statistic and its degrees of freedom, as well as its probability. Often you will get some additional statistics as well.

**Additional Analyses**

The remaining analyses were not done on this data set, but some comments about them are included for completeness.

**Simple Frequencies**

You can use Pivot Tables to get simple frequencies.  (see Crosstabulations for more about how to get Pivot Tables.)  Using Pivot Tables, each column is considered a separate variable, and labels in row 1 will appear on the output.  You can only do one variable at a time.  
   
Another possibility is to use the Frequencies function.  The main advantage of this method is that once you have defined the frequencies function for one column, you can use Copy/Paste to get it for other columns.  First, you will need to enter a column with the values you want counted (bins).  If you intend to do the frequencies for many columns, be sure to enter values for the column with the most categories.  e.g., if 3 columns have values of 1 or 2, and the fourth has values of 1,2,3,4, you will need to enter the bin values as 1,2,3,4. Now select enough empty cells in one column to store the results - 4 in this example, even if the current column only has 2 values.  Next choose Insert/Function/Statistical/Frequencies on the menu.  Fill in the input range for the first column you want to count using relative addresses (e.g. A1:A100).  Fill in the Bin Range using the absolute addresses of the locations where you entered the values to be counted (e.g. $M$1: $M$4). Click Finish. Note the box above the column headings of the sheet, where the formula is displayed.  It starts with "= FREQUENCIES (".  Place the cursor to the left of the = sign in the formula, and press Ctrl-Shift-Enter.  The frequency counts now appear in the cells you selected.

To get the frequency counts of other columns, select the cells with the frequencies in them, and choose Edit/Copy on the menu. If the next column you want to count is one column to the right of the previous one, select the cell to the right of the first frequency cell, and choose Edit/Paste (Ctrl-V).  Continue moving to the right and pasting for each column you want to count.  Each time you move one column to the right of the original frequency cells, the column to be counted is shifted right from the first column you counted.

If you want percent’s as well, you’ll have to use the Sum function to compute the sum of the frequencies, and define the formula to get the percent for one cell.  Select the cell to store the first percent, and type the formula into the formula box at the top of the sheet - e.g. = N1\*100/N$5 - where N1 is the cell with the frequency for the first category, and N5 is the cell with the sum of the frequencies.  Use Copy/Paste to get the formula for the remaining cells of the first column.  Once you have the percent’s for one column, you can Copy/Paste them to the other columns.  You’ll need to be careful about the use of relative and absolute addresses! In the example, above, we used N$5 for the denominator, so when we copy the formula down to the next frequency on the same column, it will still look for the sum in row 5; but when we copy the formula right to another column, it will shift to the frequencies in the next column.

Finally, you can use Histogram on the Data Analysis menu. You can only do one variable at a time.  As with the Frequencies function, you must enter a column with "bin" boundaries. To count the number of occurrences of 1 and 2, you need to enter 0,1,2 in three adjacent cells, and give the range of these three cells as the **Bins** on the dialog box.    The output is not labeled with any labels you may have in row 1, nor even with the column letter.   If you do frequencies on lots of variables, you will have difficulty knowing which frequency belongs to **which column of data.**

**Linear Regression**

Since regression is one of the more frequently used statistical analyses, we tried it out even though we did not do a regression analysis for this example. The Regression procedure in the Data Analysis tools lets you choose one column as the dependent variable, and a set of contiguous columns for the independents. However, it does not tolerate any empty cells anywhere in the input ranges, and you are limited to 16 independent variables. Therefore, if you have any empty cells, you will need to copy all the columns involved in the regression to new columns, and delete any rows that contain any empty cells. Large models, with more than 16 predictors, cannot be done at all.

**Analysis of Variance**

In general, the Excel's ANOVA features are limited to a few special cases rarely found outside textbooks, and require lots of data re-arrangements.

**One-way ANOVA**

Data must be arranged in **separate and adjacent** columns (or rows) for each group.  Clearly, this is not conducive to doing 1-ways on more than one grouping.  If you have labels in row 1, the output will use the labels.

**Two-Factor ANOVA Without Replication**

This only does the case with **one observation per cell** (i.e. no Within Cell error term).  The input range is a rectangular arrangement of cells, with rows representing levels of one factor, columns the levels of the other factor, and the cell contents the one value in that cell.

**Two-Factor ANOVA with Replicates**

This does a two-way ANOVA with equal cell sizes.  Input must be a rectangular region with columns representing the levels of one factor, and rows representing replicates within levels of the other factor.  The input range MUST also include an additional row at the top, and column on the left, with labels indicating the factors.  However, these labels are not used to label the resulting ANOVA table.  Click Help on the ANOVA dialog for a picture of what the input range must look like.  
  
**Requesting Many Analyses**

If you had a variety of different statistical procedures that you wanted to perform on your data, you would almost certainly find yourself doing a lot of sorting, rearranging, copying and pasting of your data. This is because each procedure requires that the data be arranged in a way, often different from the way another procedure wants the data arranged. In our small test, we had to sort the rows to do the t-test, and copy some cells to get labels for the output. We had to clear the contents of some cells to get the correct paired t-test, but did not want those cells cleared for some other test. And we were only doing five tasks. It does not get better when you try to do more. There is no single arrangement of the data that would allow you to do many different analyses without making many different copies of the data. The need to manipulate the data in many ways greatly increases the chance of introducing errors.

Using a statistical program, the data would normally be arranged with the rows representing the subjects, and the columns representing variables (as they are in our sample data). With this arrangement, you can do any of the analyses discussed here, and many others as well, without having to sort or rearrange your data in any way. Only much more complex analyses, beyond the capabilities of Excel and the scope of this article would require data rearrangement.

**Working with Many Columns**

What if your data had not 4, but 40 columns, with a mix of categorical and continuous measures? How easily do the above procedures scale to a larger problem?

At best, some of the statistical procedures can accept multiple contiguous columns for input, and interpret each column as a different measure. The descriptive and correlations procedures are of this type, so you can request descriptive statistics or correlations for many continuous variables, as long as they are entered in adjacent columns. If they are not adjacent, you need to rearrange columns or use copy and paste to make them adjacent.

Many procedures, however, can only be applied to one column at a time. T-tests (either independent or paired), simple frequency counts, the chi-square test of independence, and many other procedures are in this class. This would become a serious drawback if you had more than a handful of columns, even if you use cut and paste or macros to reduce the work. In addition to having to repeat the request many times, you must decide where to store the results of each, and make sure it is properly labeled so you can easily locate and identify each output.

Finally, Excel does not give you a log or other record to track what you have done. This can be a serious drawback if you want to be able to repeat the same (or similar) analysis in the future, or even if you’ve simply forgotten what you’ve already done.

Using a statistical package, you can request a test for as many variables as you need at once. Each one will be properly labeled and arranged in the output, so there is no confusion as to what’s what. You can also expect to get a log, and often a set of commands as well, which can be used to document your work or to repeat an analysis without having to go through all the steps again.

**Summary**

Although Excel is a fine spreadsheet, it is not a statistical data analysis package. In all fairness, it was never intended to be one. Keep in mind that the Data Analysis Tool Pak is an "add-in" - an extra feature that enables you to do a few quick calculations. So, it should not be surprising that that is just what it is good for - a few quick calculations. If you attempt to use it for more extensive analyses, you will encounter difficulties due to any or all the following limitations:

* Potential problems with analyses involving missing data. These can be insidious, in that the unwary user is unlikely to realize that anything is wrong.
* Lack of flexibility in analyses that can be done due to its expectations regarding the arrangement of data. This results in the need to cut/paste/sort/ and otherwise rearrange the data sheet in various ways, increasing the likelihood of errors.
* Output scattered in many different worksheets, or all over one worksheet, which you must take responsibility for arranging in a sensible way.
* Output may be incomplete or may not be properly labeled, increasing possibility of misidentifying output.
* Need to repeat requests for some analyses multiple times to run it for multiple variables, or to request multiple options.
* Need to do some things by defining your own functions/formulae, with its attendant risk of errors.
* No record of what you did to generate your results, making it difficult to document your analysis, or to repeat it later, should that be necessary.

If you have more than about 10 or 12 columns, and/or want to do anything beyond descriptive statistics and perhaps correlations, you should be using a statistical package. There are several suitable ones available by site license through OIT, or you can use them in any of the OIT PC labs. If you have Excel on your own PC, and don’t want to pay for a statistical program, use Excel to enter the data (with rows representing the subjects, and columns for the variables). All the mentioned statistical packages can read Excel files, so you can do the (time-consuming) data entry at home, and go to the labs to do the analysis.

Eva Goldwater  
Biostatistics Consulting Center  
University of Massachusetts School of Public Health   
updated February 2007