# Using Microsoft Excel for Probability and Statistics

## *Introduction*

Despite having been set up with the business user in mind, Microsoft Excel is rather poor at handling precisely those aspects of statistics which might be most useful in the business context, namely informative displays of data. Calculations, on the other hand, can be done quite quickly and (in most cases) accurately in Excel. An additional use of Excel is as a generator of statistical tables: there are built-in routines to calculate percentage points or p-values for many distributions of interest in statistical modelling. The use of Excel to simulate random variables will also be discussed.

### 1. Displaying Data

Chart Wizard is a splendid tool for producing colourful, eye-catching graphics. The problem with Chart Wizard is that impressive effects draw attention away from the contents of the graph: if you want your readers to look at what the graph is saying, you need to keep the diagram simple.

Excel is good at scatter diagrams, useful for Linear Regression. Select a block of cells two columns wide then use the Chart Wizard (the toolbar icon is a multi-coloured 3-dimensional bar chart) to choose XY (Scatter) chart type, avoiding the ones which join up the dots. The default assumption is that the first column contains the *x*-data, the second the *y*-data, so I recommend that you arrange your data that way in the first place. Once the chart is in place you can add the line of best fit by right-clicking on one of the points on the chart and selecting "Add Trendline..." from the menu which appears. To display the equation of the line and the $R^2$ value you need to select the Options tab.

As far as histograms, stem-and-leaf diagrams, box-and-whisker plots, normal probability plots and others are concerned, Excel is a bit of a washout. You can get Excel to produce a reasonable cumulative frequency diagram, but only by calculating the *y*-values yourself and telling Excel that you want an XY (Scatter) chart with lines joining the points. Minitab is much better than Excel at graphical procedures.

### 2 Entering Data

Although it is possible to use Excel when data are arranged in rows, it is preferable to arrange the data in such a way that each sample occupies a column.

*A lecturer has two groups of students doing the same course. The coursework marks of the students in the first group are 11, 15, 18, 20, 17, 20, 14 and 16. The second group have scored 14, 19, 17, 13 and 9. The maximum possible mark for the coursework is 20.*

Type in the words `Coursework Marks` in cell A1 as the title of the sheet. Type in `Group 1` in cell B3, `Group 2` in C3, and enter the data in the appropriate columns, starting from row 4.

### 3 Introduction to worksheet functions

The statistical worksheet functions in Excel almost all work on ranges. The range can be entered either directly, for example A5:A24 (the cells in column A from row 5 to row 24 inclusive: note the use of a colon, not a comma), or else by giving the range a name (using Insert | Name | Define) and using that name. It is usually preferable to use names when you are going to refer to a sample more than once.

Select cells B4:B11, choose Insert | Name | Define and accept the default name Group_1. Similarly use the name Group_2 for the data in C4:C8.

The first idea that occurs to us might be to calculate the means of the two samples. The Excel function which calculates sample means is called `AVERAGE().` (In Excel you can always use lower case letters instead of upper case for function names and cell references; I just use upper case to make it stand out in the printed notes.)

In cell A13 type the words `Sample mean` as a label, then in B13 enter =`AVERAGE(Group_1).` You should find that the average of the values in column B is now visible in cell B13 in the main body of the spreadsheet.

Find the sample mean of Group 2's scores similarly.

A general point about worksheet functions like `AVERAGE()` is that you must not leave a space between the name of the function and the opening parenthesis, otherwise Excel will consider it an error.

## 4 Relative and Absolute Addressing

Copying the formula =`AVERAGE(Group_1)` from cell B13 into cell C13 is not helpful, as it leaves you with the average of Group 1 rather than Group 2. But if you had typed =`AVERAGE(B4:B11)` into B13 and then copied that into C13, you would have found that it gave you the right result. Why are the results different?

The difference lies in the fact that B4:B11 is a relative address, whereas Group_1, which could equally well be written $B$4:$B$11, is an absolute address. When a formula involving a relative address is copied into a cell which is one column across and 3 rows down from its original position, the formula in the new cell refers to a range which is shifted by one column and 3 rows from the original range. A formula involving an absolute address always refers to the same address, even if the formula is copied to another cell.

Investigate the effects of the half-absolute addresses $B4:$B11 and B$4:B$11 by using them, one at a time, in the AVERAGE formula and copying the formula across to other cells.

See Contingency Tables in Section 10 for an application which shows their importance.

## 5 More Worksheet Functions for Statistics

`AVERAGE()` is only one of a large number of functions which Excel provides to assist with the manipulation of numbers. It is often difficult to remember the precise name of a particular function, or the order of the parameters. To assist you in this, Excel provides the function wizard.

If you press the button on the toolbar labelled ***f**x*, this invokes the function wizard, which starts off by trying to determine which function you need. As you use the program you will find the functions that you call upon are kept in the function category called Most Recently Used, but to start with you will have to find the functions elsewhere. Most, though not all, of the functions we use here will be in the Statistical category.

Practice with the function wizard. In rows 14 to 19 of column A type in text for labels: Standard Deviation, Maximum, Upper Quartile, Median, Lower Quartile and Minimum. In columns B and C enter formulas to calculate the required values: the functions you will need are `STDEV`, `MIN`, `MAX`, `MEDIAN` and `QUARTILE.` For the first four you can type the formula directly into the cell, not forgetting to begin the formula with an = sign, or you can use function wizard to guide you through. When it comes to quartiles you are probably better off using the function wizard, as the `QUARTILE` function needs to know not only the range but also which quartile is required (1 for lower, or 3 for upper).

`MIN`, `MAX` and `MEDIAN`, like `AVERAGE`, can take more than one argument. In this instance we only enter the range `Group_1` (or `Group_2`) as the first argument and leave all the others blank, but it is no harder to calculate the sample statistics for a set of values which do not fill a rectangular area of the sheet.

Other simple functions include:
`COUNT(range)` – the number of cells in the range which contain numerical data
`VAR(range)` – the sample variance of the range
`SUM(range)` – the sum of the values in the range
`SUMSQ(range)` – the sum of the squares of the values in the range
`DEVSQ(range)` – $\sum (x_i - \bar{x})^2$

## 6 Functions useful for calculating probabilities

As a short-cut to calculating probabilities for standard distributions it is possible to use:
`BINOMDIST(x,k,θ,0)` – this is $P(X = x)$, i.e. the probability function, when $X \sim \text{Bin}(k, \theta)$
`BINOMDIST(x,k,θ,1)` – this is $P(X \leq x)$, i.e. the distribution function, when $X \sim \text{Bin}(k, \theta)$
`POISSON(x,μ,0)` and `POISSON(x,μ,1)` – similarly, when $X \sim \text{Poisson}(\mu)$
`NEGBINOMDIST(x,1,θ)` – the probability function of Geometric(θ)
`NORMSDIST(z)` – $P(Z \leq z)$ when $Z \sim N(0,1)$, i.e. the standard Normal distribution function $\Phi$.
`NORMDIST(x,μ,σ,1)` – $P(X \leq x)$ when $X \sim N(\mu,\sigma^2)$
`NORMDIST(x,μ,σ,0)` – the density function of $X$ when $X \sim N(\mu,\sigma^2)$
`EXPONDIST(x,λ,1)` and `EXPONDIST(x,λ,0)` – similarly for the case where $X \sim \text{expo}(\lambda)$

These probabilities are also helpful when you have to calculate p-values for non-Normal distributions. For example, the p-value associated with the observed value $T_{obs}$ of the t-statistic for a two-sided t test with 24 degrees of freedom is given by `TDIST`($T_{obs}$, 24, 2). Functions `FDIST` and `CHIDIST` are also available, for the F and $\chi^2$ distributions, but should be treated with caution because the command syntax is not the same as for the `TDIST` command.

## 7 Functions useful for calculating critical values

Percentage points of all the standard distributions are available, but be careful, because there is no uniformity of syntax. In all these examples the 5% point is given; for different percentage points it should be elementary to adjust the function appropriately.

`NORMSINV(0.95)` gives the 5% point of the standard N(0,1) distribution.

`NORMINV(0.95,μ,σ)` – the same, but for $N(\mu, \sigma^2)$

`TINV(0.1,ν)` – note that TINV is intended for use in calculating two-sided confidence intervals, which is why the first argument is the confidence level (10%) instead of the percentage point (5%).

`FINV(0.05, ν1, ν2)` – the (upper) 5% point of the F distribution with numerator degrees of freedom equal to $\nu_1$ and denominator df $\nu_2$. For the lower 5% point you can use either `FINV(0.95, ν1, ν2)` or `1/FINV(0.05, ν2, ν1)` (note the interchange of the df parameters here).

`CHIINV(0.05,ν)` – again the upper 5% point; use `CHIINV(0.95,ν)` for the lower 5% point.

## 8 Functions for hypothesis testing and confidence intervals

Excel provides:

`ZTEST(sample, μ0, σ)` – the p-value associated with a test of $H_0$: $\mu = \mu_0$ against $H_1$: $\mu \neq \mu_0$ when the standard deviation of the population is known to be $\sigma$.

`TTEST(xsample, ysample, 2, 2)` – the p-value associated with a test of $H_0$: $\mu_X = \mu_Y$ against $H_1$: $\mu_X \neq \mu_Y$ using the standard two-sample t test. The third parameter controls whether the test is a one-sided or two-sided test, the final parameter may be adjusted to give a paired-sample test or the Welch test (an approximate t test in the case where equal variances are not assumed).

`CHITEST`(observed_range,expected_range) – the p-value associated with the contingency table procedure. You need to work out the expected values and arrange them in a rectangle of the same dimensions as the rectangle of observed values.

## 9 Functions useful for linear regression

The least squares estimates $a$ and $b$ of the parameters $\alpha$ and $\beta$ of the linear regression model $Y = \alpha + \beta x + E$ can be obtained as $a$ = `INTERCEPT`(*y-range*, *x-range*), $b$ = `SLOPE`(*y-range*, *x-range*).

Quantities used in the analysis of a regression can be obtained as $S_{xx}$ = `DEVSQ`(*x-range*), $S_{yy}$ = `DEVSQ`(*y-range*), $S_{xy} = b * S_{xx}$, $\hat{s}^2 = S_{yy} - b * S_{xy}$.

Fitted values are calculated as $a + b*x$, or alternatively using TREND(y-range, x-range, x-range); residuals are given by Observed – Fitted.

`FORECAST`($x_0$, y-range, x-range) calculates $a + b*x_0$ for the purpose of forecasting future *y*-values.

Also useful are `CORREL`(*y-range*, *x-range*), the correlation coefficient, and `RSQ`(*y-range*, *x-range*), the value of $R^2$. You might also care to investigate `STEYX`(*y-range*, *x-range*), which claims to give estimates of the standard error of each *y*-value in the regression.

## 10 Ranks

If you need to conduct a test based on ranks, it would of course be helpful to have a reliable function to calculate the relative ranks of the observations. Excel provides a function which is not quite what is needed, so a certain amount of extra effort is required. The problem lies in the way Excel handles ties: for example, if the highest value in the sample is achieved by three observations, they will each be given rank 1 instead of (as required by rank-based methods) rank 2.

Some thought will enable you to get around this problem. It is clear that we need to add to the rank calculated by Excel a quantity which is 0.5 * (number of observations which have this value – 1). So the formula to calculate the rank of one value, say A20, in a sample, taking proper account of ties, is

```
RANK(A20, WholeSample) + 0.5*(COUNTIF(WholeSample, "=" & A20)-1)
```

Here the `COUNTIF()` function is used to count the number of values in the whole sample which satisfy the condition that they are equal to A20.

If you prefer to have your numbers ranked in ascending order instead, use

```
RANK(A20,WholeSample,1) - 0.5*(COUNTIF(WholeSample,"="&A20)-1)
```

## 11 Goodness-of-Fit tests

The procedure for Goodness-of-Fit tests is to arrange the Observed frequencies in one column the Expected frequencies in a separate column the same size and shape, and the values of $(O_i - E_i)^2 / E_i$ in yet a third column. These last can then be summed to give the $X^2$ value.

To calculate the Expected values, you will first need to estimate the parameter of the distribution, then to use the estimate to calculate probabilities. For example, when testing the fit of a Poisson distribution, first find the sample mean of the data to estimate the parameter of the Poisson, then calculate expected frequencies using the formula `=N*POISSON(i,LambdaEst,0)`, where *N* is the total number of observations, perhaps obtained as =`SUM(frequencies).`

The procedure is more difficult for continuous distributions, but only because you have to choose the end-points of the intervals into which the observations are sorted. For example, suppose you want to test normality based on a division of the sample into the intervals (–∞, 1), (1,2), (2,3), (3,5) and (5,∞). First estimate μ and σ in the usual way, name the estimators MuHat and SigmaHat, then calculate

```
N* NORMDIST(1, MuHat, SigmaHat, 1)
N*(NORMDIST(2,MuHat,SigmaHat,1)-NORMDIST(1,MuHat,SigmaHat,1)), etc.
```

## 12 Contingency Tables

The Observed and Expected values will each occupy a rectangle rather than a column. Suppose the Observed area is A4:C8. On the table of Observed values, use SUM() to calculate row sums and column sums, entering the values at the end of the row (D4:D8) or foot of the column (A9:C9). Find the overall sum of the observed frequencies similarly. With a little care we can fill in the whole Expected table with a single formula copied into all cells. In F4 type `=$D4*A$9/$D$9` and copy it into F4:H8.

You can now use the `CHITEST()` function to calculate the p-value of the $X^2$ statistic.

## 13 The Analysis TookPak

Excel provides a number of statistical features by means of an Add-In known as the Analysis ToolPak. If this is already set up for your version of Excel you will see the item Data Analysis... appearing at the bottom of the Tools menu. If not, choose Tools | Add-Ins... and select both Analysis ToolPak and Analysis ToolPak VBA.

The routines provided by the Analysis ToolPak are generally adequate, but have the disadvantage that they are not linked to the data: unlike the worksheet functions, they will not update themselves automatically if you change the original data.

**Descriptive statistics**: this tool provides a collection of sample statistics for a given range of data, including sample mean, variance, median, quartiles, mode and a number of others.

**F-Test Two-Sample for Variances**: tell it where to find the two samples and this procedure will give you a one-sided p-value (double it to get the usual two-sided p-value) and will tell you the cut-off point for significance at whatever level you decide is suitable.

**Regression**: gives estimates of the parameters α and β, reports $R^2$ and adjusted $R^2$, gives the estimate of σ; will report fitted values and residuals if required. Summarises the statistical findings in two tables: one gives estimated values of parameters and confidence intervals, the other decomposes the total variation into residual sum of squares and error sum of squares.

**t-Test: Paired Two Sample for Means**: assumes that the two samples you give it are Before and After samples, then gives p-values for one-sided and two-sided tests based on the t distribution.

**t-Test: Two-Sample Assuming Equal Variances**: the standard two-sample t-test as performed in lectures.

**z-Test: Two Sample for Means**: this tests for the equality of the means of two samples if the variance of each sample is assumed known (the variances need not be equal).

## 10 Simulation

Excel will generate uniform random numbers in the range from 0 to 1 using the `RAND()` function. This function is slightly disconcerting, as the random numbers are generated all over again every time you require the worksheet to perform any calculation at all.

Although the theory of the simulation of random variables does not come until Part II, you may be interested to know that the inverse distribution function turns uniform random variables into observations from the distribution of interest. Thus

`NORMSINV(RAND())` generates a simulated N(0,1) observation;
`NORMINV(RAND(),Mu,Sigma)` generates a simulated $N(\mu,\sigma^2)$ observation;
`CHIINV(RAND(), 10)` generates a simulated $?^{2}_{10}$ observation.

Unfortunately the theory is not quite so simple for discrete random variables and it requires a little more effort to simulate Binomial or Poisson. On the other hand, it is quite possible to simulate a Bernoulli random variable with parameter $p$: the formula `IF(RAND()<p,1,0)` will do that. Adding together 20 Bernoulli($p$) random variables gives Binomial(20,$p$). It may not be the most efficient way to simulate Binomials, but it works.

**EXERCISE (Optional. Anyone can do it; the mark obtained will replace your lowest quiz mark)**

Download the data from the website and paste them into Excel. Note that the data are generated randomly and independently for each student.

i) Calculate a 95% confidence interval for the male's mean terrorimeter reading when confronted by a random Martian. Store the lower end in F3, upper end in G3.

ii) Carry out an F test to determine whether the variances of the male's and female's terrorimeter readings can be assumed equal. Store the F statistic in F5, the p-value in G5.

iii) Find the line of best fit for the regression of the female's terrorimeter readings against the Martians' Fang Factor. Store $a$ in F7, $b$ in G7.

iii) Test whether the male's terrorimeter readings are exponentially distributed. Store the $X^2$ statistic in F9, the degrees of freedom in G9 and the p-value in H9.

All calculations must use Excel worksheet functions. Arrange your workings on the worksheet so that they print out on a single sheet of paper. (Use Print Preview to check.) Put your name and the name of your statistics tutor at the top and print off a copy. Now go to Tools | Options | View and check the box labelled Formulas. Your worksheet will show all the formulas you used to obtain your results. Print off a copy of this as well. Submit both sheets, stapled together, to Russell Gerrard by **Friday 30 April**.