Doing Statistics with Excel

Lecture 51

Robb T. Koether

Hampden-Sydney College

Fri, Apr 25, 2008

Outline

1 Introduction

2 Cells, Values, and Formulas

3 Functions

4 Statistical Functions

5 The Regression Line

6 Testing the Significance of the Regression Line

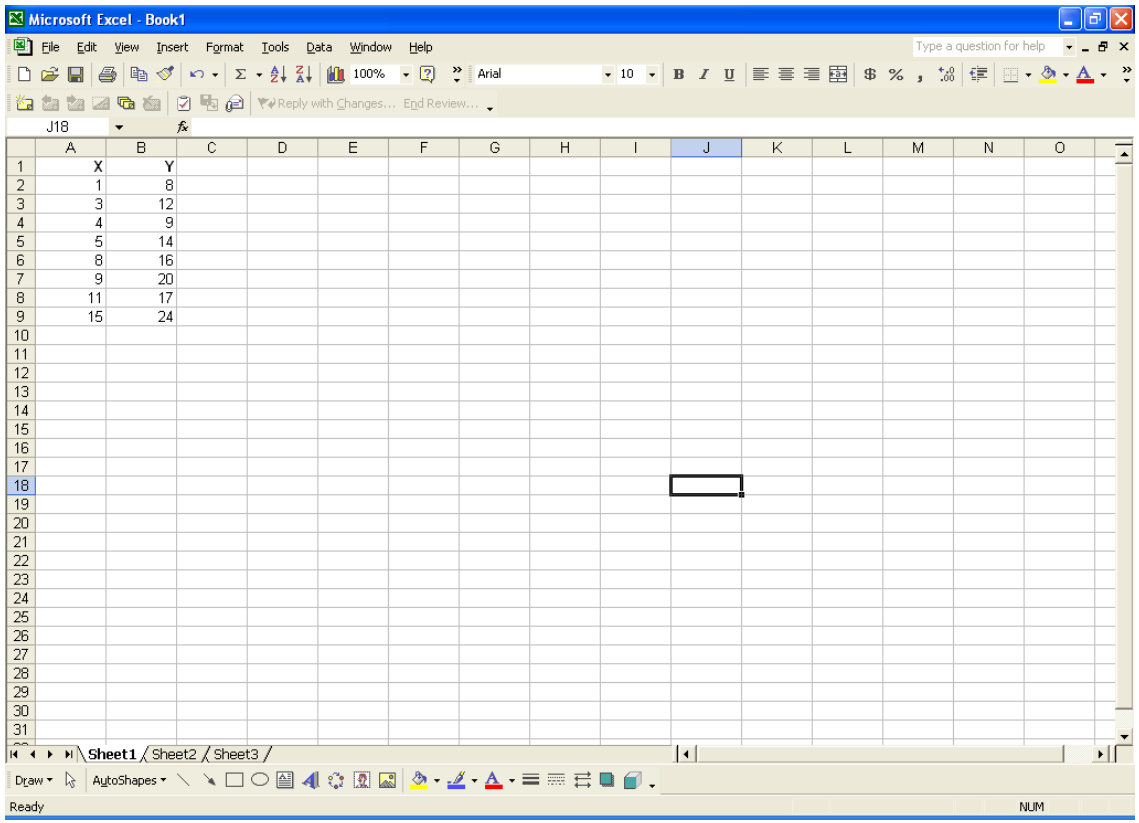
Introduction

- The spreadsheet Excel has many statistical functions.
- Most of them are easy to use, but one must understand the statistics behind them to know how and when to use them.
- In this lecture, we will look at a few functions that perform tasks with which we are familiar.

Cells, Values, and Formulas in Excel

- The spreadsheet is arranged as an array of rows and columns.
- The rows are numbered `1`, `2`, `3`, ...
- The columns are labeled `A`, `B`, `C`, ...
- An individual cell is referenced by its column label and row number, e.g., `C5`.

Cells, Values, and Formulas in Excel



	A	B
1	X	Y
2	1	8
3	3	12
4	4	9
5	5	14
6	8	16
7	9	20
8	11	17
9	15	24

Cells, Values, and Formulas in Excel

- One cell may reference another cell.
- Begin with an equal sign `=`.
- Then type an expression to be calculated.
- For example, in cell `C1`, we may write `=2*B1+10`.

Using Functions in Excel

- There is a variety of mathematical functions that can be included in a formula.
- Type = and then select `Insert > Function...` from the menu bar. A box opens.
- Choose the category of function, then choose the function.
- For example, in the `Math & Trig` category, we could choose `SQRT()`, `RAND()`, or `SUM()`.

Using Functions in Excel

- Some functions apply to a range of cells, e.g., the `SUM()` function.
- We could write `=SUM(A1:A8)` to get the total of cells `A1` through `A8`.
- Or we could write `=SUM(A1:B8)` to get the total of cells `A1` through `A8` plus `B1` through `B8`.

Fill Right and Fill Down

- Typically, if we want to fill a formula in every cell in a range, we can use the “fill right” or “fill down” features.
- To fill right,
  - Enter the formula in the leftmost cell of the range.
  - Select the range of cells.
  - Press `CTRL-R`.
- To fill down,
  - Enter the formula in the uppermost cell of the range.
  - Select the range of cells.
  - Press `CTRL-D`.

Statistical Functions

- Now let's look at the Statistical functions.
- The `AVERAGE()` function will compute the average of the given cells.
- There are several functions for the standard deviation.
  - `STDEV()` computes the standard deviation s of a sample.
  - `STDEVP()` computes the standard deviation σ of a population.
- There are many others, for example,
  - `MEDIAN()` computes the median of a range.
  - `QUARTILE()` computes a quartile.
  - `PERCENTILE()` computes a percentile.

Excel and Various Sums of Squares

- There are various math functions that will compute sums and sums of squares
  - `SUM()` $= \sum x$.
  - `SUMSQ()` $= \sum x^2$.
  - `SUMX2MY2()` $= \sum (x^2 - y^2)$.
  - `SUMXMY2()` $= \sum (x - y)^2$.

Excel and Various Sums of Squares

- There is also a statistical function that will compute the sum of squared deviations SSX.
  - `DEVSQ()` $= \sum (x - \overline{x})^2$

Excel and the Correlation Coefficient

- We can find the correlation coefficient by using the various summing functions.
- Or we can use the function `CORREL()`.

Excel and the Regression Line

- We can obtain a regression line graphically by first creating a scatterplot (details omitted).
- Then select `Add Trendline...` (details omitted).
- Select `Linear` and the regression line appears (details omitted).
- Also select
  - `Display equation`
  - `Display R-squared value`

  to see the equation of the regression line and the value of r^2.

Excel and the Regression Line

- We can obtain the regression coefficients a and b within cells by using statistical functions.
  - `SLOPE()` computes the slope of the regression line.
  - `INTERCEPT()` computes the y-intercept of the regression line.
- The function `TREND()` will use the regression line to predict y for a given value of x.

Testing the Significance of β and ρ

- By doing the necessary calculations, we can perform the linear regression t test to determine whether β and ρ are statistically significant.
- To do this, we need one more function.
  - `TDIST()` computes the area of a tail (or two tails) of the t distribution with the given number of degrees of freedom.