

NOTE G: DATA INPUT ERRORS

Spreadsheet setup and data entry may at times create hidden or inherent faults and errors. These are not detected, and carry over into faulty or erroneous outputs. This has to do with the way the external/internal world is described by using symbols, language and numbers. It also occurs from the way that Excel as a spreadsheet works with data.

In using a software program, information about the external world is described as data, and represented by symbols, character strings and numbers. Excel makes certain interpretations about the data being entered. The combination of the nature of the data being entered, and the interpretation of that data by Excel create the hidden errors.

Measurement theory has identified several characteristics of data described by means of numbers. A collection of digits can represent different levels of measurement, which have different properties. These levels are nominal, ordinal, interval, ratio and absolute. (See Sarle 1996 for definitions and characteristics of each of these.) A number can be assigned to different objects at the nominal level, but numerical operations such as summing, differencing, multiplying and dividing on the numbers have no meaning.

Objects can be assigned numbers or text strings. Categorical data is an example of the application of text strings. Categorical data represents a measurement from a categorical treatment, where the categories are descriptions (text) grouped into classes, which may or may not have numbers associated with the classes. Since the independent variables are non-numeric, they are treated here as a separate level of measurement. Dichotomous variables are a form of categorical data in which there are only two levels, in a bin or not in a bin. If there are to be arithmetic operations (i.e. add, subtract, multiply and divide) on the entered digits, then the data should have a certain level of measurement in order for the results to be valid. Below is a brief table giving allowable statistical calculations on five types.

Table F-1: Use of Functions on Data With Different Levels of Measurement

Function/Operation	Nominal	Ordinal	Interval	Ratio	Categorical
Average			yes	yes	
Geometric & Harmonic Means				yes	
Standard Deviation/Variance			yes	yes	
Rank		yes	yes	yes	
Median		yes*	yes	yes	
Mode			yes	yes	
Regression			yes	yes	
ANOVA		yes	yes	yes	yes
Equal to or not equal to	yes	yes	yes	yes	yes
Greater than or less than		yes	yes	yes	

*Only an odd number of values

The absolute level of measurement represents “a count of” and is inherently integer only data. All the function/operations of the ratio type can be done on data with an absolute

level of measurement. The resulting decimal outputs may have to be converted to integers, since decimal or fractional values do not logically exist. This presents a problem with quartiles and the median, which is discussed in note B of appendix B. This presents problems for computations of degrees of freedom (df) values where two or more data sets are combined for a statistical analysis. This is discussed under the Fisher-Berens problem in appendix B. This problem also occurs when the inverse of the BINDIST function is called.

For example, if the data being analyzed includes a variable that has a level of measurement that is ordinal or between ordinal and interval, then the results of a regression with that variable is in error. The standard error and the coefficient values from the regression are wrong, because of the uncertainty about the nature of the intervals between the numbers entered as data. A lot of market research and survey data falls into this category. The usual assumption is to just overlook or ignore this problem because the interval widths cannot be measured.

The approach to putting data into a blank cell requires the software to somehow detect entry and determine what the data is from many possible types of data being entered. Microsoft created a data type called “variant” which contains the data entered into each cell. For a worksheet of 256 columns by 65,536 rows (maximum allowable), there are 16,777,216 cells or variant variables that can potentially be assigned to that worksheet (if completely filled) (However there are internal limitations on the size.)

A variant can hold different types of data. The variant type does not track the level of measurement of the number or text. The baseline is that every number entered is treated as being from a ratio level of measurement. The variant length is 16 bytes (128-bits) for non-text data and 26 bytes (208-bits) for variable length strings. Sixteen bits of the length are used for identification of the data type. Excel treats entered integers as floating point numbers.

Table F-2: Data That Can Be Held Within A Variant Variable

VarT type Value	Name	Contents of Variant
0	Empty	There is no data in the variant
1	Null	The variant has no value, different from being empty
2	Integer	A whole number between -32,768 and 32,767
3	Long	A whole number between -2,147,483,648 and 2,147,483,647
4	Single	A floating point number of 32 bits. +3.402823E+38 to +1.401298E-45 or -1.401298E-45 to -3.402823E+38
5	Double	A floating point number of 64 bits. +1.707693134862315D+308 to +4.94066458412465D-324 or -4.940656458412465D-324 to -1.797693134862315D+308
6	Currency	A decimal number with 4 fixed decimal places. -922,337,203,685,477.5808 to 922,337,203,685,477.5807

VarT ype Value	Name	Contents of Variant
7	Date/Time ¹	A number representing a combination of the date (left) and time (right)
8	String	Text
9	OLE object	Object
10	Error	Error codes
11	Boolean	Binary, true or false
12	Variant	An array of variants
13	Non-OLE object	Object
14	Decimal	A decimal number of 28 digits with a floating decimal point
17	Byte	Byte
8192	Array	An ordered table of values. A custom data type

The function VARTYPE can identify under VBA, in a macro which of the above the variant is. At the worksheet/cell level the TYPE(...cell being referenced...) function will return 1 if the cell being referenced is a number (double), 2 if it is text, 4 if it is logical, 16 if it contains an error value and 64 if it is a reference to an array of variants.

The user has no direct control on establishing what type of data the contents of a cell represents. Excel internally determines the data type based on the data being entered. Each cell is treated as an object with many properties, among which are the format properties of the cell. This is called up when the menu sequence (Format>Cells>Number) is clicked.

General (The default, uses Excel internal logic to determine type of data entered)

Number (Sets in the double type)

Currency (Changes the double type to the currency type)

Accounting (Sets in the double type)

Date (Sets in the Date/Time type, outputs the left value as a date)

Time (Sets in the Date/Time type, outputs the right value as a time)

Percentage (Sets in the double type, with percentage formatting)

Fraction (Sets in the double time with fraction formatting)

Scientific (Sets in the double type with scientific notation outputs)

¹ Excel assigns serial values to days, hours, minutes and seconds in a long integer format which allows for date/time arithmetic. The default basic unit is days starting from Sunday January 1, 1900. The date is a serial number of days from this date. The limit is December 31, 2029 if only a 2 digit (i.e. 12/20/05) year is entered, since Excel assumes 2000 as the unspecified year. One should always use a 4 digit year to avoid interpretation errors.

Text (Sets in the string type. Text format cells are treated as text even when a number is in the cell. The cell is displayed exactly as entered.)

Special (Sets in the double type, with special output formats)

Custom (User selects/modifies standard format property coding)

Hidden errors can occur when data are entered. For example with the *General* format on each cell (the default), you enter the following:

A number (digits) {Enter}. The variant type is set as double.

A number with digits and accidental spaces {Enter}. The variant type is set as double.

A number with beginning quotes {Enter}. The variant type is set as text.

A number with an accidental character {Enter}. The variant type is set as text. You recognize the error and edit the character out in the formula box. The variant type remains as text.

Sometimes Excel logic will change the variant type to double and sometimes not. In Excel 97 this was a problem, but appears to have been fixed in Excel 2000. **In Excel 2003, there is a tendency to lock into text if an error is made. It cannot be changed by re-entry. One has to select: Edit-Clear-All when the cell is selected, to clear out the lock.**

If the cell is initially empty and the format general, then Excel correctly sets the data type. If the cell had an error and is being changed, Excel may not correctly reset the type. If not changed, the hidden error is now in the cell. If later on, this list is a range, and the range is input to STDEV, the output from STDEV will be in error because the cell (d.) was identified as being text and was skipped in the calculation, and there was no evidence that this error has occurred.

The only fix is to show when this number/text error occurs.

One way, is to set up the cell format as

Format>Cells

Number → General

Alignment/Horizontal → General

Expand column width

When the cell is number (Double), the digits are right side anchored.

When the cell is text (String), the characters are left side anchored.

When the cell is logical (Boolean), the words TRUE and FALSE are capitalized and centered

A visual inspection down the column should show any discrepancies.

Another way is to add a new (blank) column to the right of the column having the entered data and format the column for "Number → 0 decimals". For illustration the column with the data will be assumed to be B, the data starts at row 1 and the new column is C. Put in the first cell (C1) the formula =ISNUMBER(B1) and function copy downwards to the end of the data. Column C will have 1's and 0's. 1's corresponds to numbers and 0's are

non-numbers. The function =TYPE(B1) can also be used to give numbers identifying the type of variant.

The second method takes more time, but it identifies both text and logic values positively.

There are functions where text and logic values in cells are ignored and functions where text and logic values in cells have an effect. These are the functions with an added "A" at the end of the name. In these functions, text is given a value of zero and counted in. A logic true (either "true" or 1) is given a value of 1 and counted in. A logic false (either "false" or 0) is given a value of 0 and counted in. Consequently STDEV and STDEVA may give different values from the same range.