ABSTRACT
Microsoft Excel is not regarded as a package to do statistical analysis for a number of reasons, namely a spreadsheet is not technically advanced enough to ensure proper data/documentation management and workflows, and statistical functions are lacking or not beyond the capabilities of serious statistical computational needs. Despite this, Excel is still the leading software package used to do statistics. This paper looks at two distinct table types, the Demographics and the Adverse Event tables; and three distinct categories of data, the categorical, continuous and occurrence categories, and shows that Excel with VBA can indeed be used to produce these tables. While not used for publication or serious reporting, these could be used as a way to look at the data before 'official' reports are available. Along the way, we will also see how an Adverse Event Timeline Plot may be created.

INTRODUCTION
Why use Excel? It is not regarded as a serious package to do statistical analysis. But it is easy to use, most people have access to it, it easier to get a person to do a simple Demographics table using Excel than find a person who knows SAS or R to do the same. However, Excel has one big disadvantage, managing data – getting an updated set of data for a study and rerunning SAS programs to produce output is far quicker and easier than importing data into Excel sheets and processing results.

That being said, Excel does have an advantage of seeing results in real time, the intimacy of reviewing numbers on a screen without switching between files.

This paper shows how it is possible to use Excel to produce output using different techniques which are given in the paper.

THE DATA
Two datasets will be used in this paper, ADSL:

<table>
<thead>
<tr>
<th>STUDYID</th>
<th>USUBJID</th>
<th>SUBJID</th>
<th>SAFFL</th>
<th>AGE</th>
<th>AGEU</th>
<th>SEX</th>
<th>RACE</th>
<th>ETHNIC</th>
<th>ARMCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NewSauceA NewSauceA-101</td>
<td>101 Y</td>
<td>23 YEARS</td>
<td>M</td>
<td>OTHER</td>
<td>NOT HISPANIC OR LATINO</td>
<td>SAUCE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>NewSauceA NewSauceA-102</td>
<td>102 Y</td>
<td>73 YEARS</td>
<td>M</td>
<td>WHITE</td>
<td>NOT HISPANIC OR LATINO</td>
<td>PLACEBO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>NewSauceA NewSauceA-103</td>
<td>103 Y</td>
<td>32 YEARS</td>
<td>F</td>
<td>MULTIPLE</td>
<td>NOT HISPANIC OR LATINO</td>
<td>PLACEBO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>NewSauceA NewSauceA-201</td>
<td>201 Y</td>
<td>31 YEARS</td>
<td>F</td>
<td>BLACK OR AFRICAN AMERICAN</td>
<td>NOT HISPANIC OR LATINO</td>
<td>PLACEBO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>NewSauceA NewSauceA-202</td>
<td>202 N</td>
<td>86 YEARS</td>
<td>M</td>
<td>WHITE</td>
<td>NOT HISPANIC OR LATINO</td>
<td>SCRNFAL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>NewSauceA NewSauceA-203</td>
<td>203 Y</td>
<td>49 YEARS</td>
<td>M</td>
<td>ASIAN</td>
<td>NOT HISPANIC OR LATINO</td>
<td>SAUCE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>NewSauceA NewSauceA-204</td>
<td>204 Y</td>
<td>58 YEARS</td>
<td>F</td>
<td>WHITE</td>
<td>NOT HISPANIC OR LATINO</td>
<td>PLACEBO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>NewSauceA NewSauceA-301</td>
<td>301 Y</td>
<td>51 YEARS</td>
<td>F</td>
<td>MULTIPLE</td>
<td>NOT HISPANIC OR LATINO</td>
<td>PLACEBO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>NewSauceA NewSauceA-302</td>
<td>302 Y</td>
<td>36 YEARS</td>
<td>M</td>
<td>BLACK OR AFRICAN AMERICAN</td>
<td>NOT HISPANIC OR LATINO</td>
<td>PLACEBO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>NewSauceA NewSauceA-303</td>
<td>303 Y</td>
<td>65 YEARS</td>
<td>M</td>
<td>WHITE</td>
<td>NOT HISPANIC OR LATINO</td>
<td>SAUCE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and ADAE:
**NAMING RANGES**

Excel has the ability to set a range of cells to a “Named Range”. In Excel 365, first highlight the range then put a name in the left box immediately under the ribbon, as the following example demonstrates:

Any named range must be unique, i.e. only one ‘AGE’, within each workbook.

**SUBSETTING CAPABILITY WITH SOME FUNCTIONS**

Some functions in Excel have subsetting capability, but not all. An example function call is:

```
MEDIAN(IF(SAFFL&ARMCD="Y"&UPPER(G$2),AGE))
```

where we get the MEDIAN value of named range AGE on the condition that the SAFFL=“Y” and ARMCD has the value of G$2, in this case it is either ‘Sauce’ or ‘Placebo’. The way this works is confusing at first but essentially it is the values of SAFFL and ARMCD are concatenated and compared with concatenated values of “Y” and uppercase values of cell G$2, then uses matched rows to call function MEDIAN of values for named range AGE.
THE DEMOGRAPHICS TABLE

Below is the mock Demographics table that is going to be filled:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sauce (N=xx)</th>
<th>Placebo (N=xx)</th>
<th>Total (N=xx)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years) [a]</td>
<td>xx</td>
<td>xx</td>
<td>xx</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>xx.xx (xx.xx)</td>
<td>xx.xx (xx.xx)</td>
<td>xx.xx (xx.xx)</td>
</tr>
<tr>
<td>Median</td>
<td>xx</td>
<td>xx</td>
<td>xx</td>
</tr>
<tr>
<td>Min</td>
<td>xx</td>
<td>xx</td>
<td>xx</td>
</tr>
<tr>
<td>Max</td>
<td>xx</td>
<td>xx</td>
<td>xx</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td>xx</td>
<td>xx xx</td>
</tr>
</tbody>
</table>
| Male              | xx.xx (xx.% | xx.xx (xx.% | xx.xx (xx.% |)
| Female            | xx.xx (xx.% | xx.xx (xx.% | xx.xx (xx.% |)
| Missing           | xx           | xx             | xx           |

[a] Age in years at Consent.

Text on row 2 and 4 are just that, text. As for the calculations these are:

G5: =COUNT(IF(SAFFL&ARMCD="Y"&UPPER(G$2),SUBJID))
H5: =COUNT(IF(SAFFL&ARMCD="Y"&UPPER(H$2),SUBJID))
I5: =SUM(G5:H5)

For Cell I5 the SUM function has been used to produce the total rather than calling another COUNT.

To include these in the header row for the table:

C6: =CONCAT("Sauce (N=",G5,")")
D6: =CONCAT("Placebo (N=",H5,")")
E6: =CONCAT("Total (N=",I5,")")

and the result is:
The calculations for AGE in column C are:

\[
\begin{align*}
C9 &= \text{TEXT} \left( \text{COUNT} \left( \text{IF} \left( \text{SAFFL} & \text{ARMCD}=\text{"Y"} \& \text{UPPER} \left( \text{G$2} \right) \right) \right), \text{"0"} \right) \\
C10 &= \text{CONCAT} \left( \text{TEXT} \left( \text{AVERAGE} \left( \text{IF} \left( \text{SAFFL} & \text{ARMCD}=\text{"Y"} \& \text{UPPER} \left( \text{G$2} \right) \right) \right), \text{"0.0"} \right), \text{"\,\text{\textbf{(}}\text{\textbf{)}}\text{\textbf{)}}}\right) \\
\text{CED} &= \text{TEXT} \left( \text{STDEV.S} \left( \text{IF} \left( \text{SAFFL} & \text{ARMCD}=\text{"Y"} \& \text{UPPER} \left( \text{G$2} \right) \right) \right), \text{"0.00"} \right) \\
C11 &= \text{TEXT} \left( \text{MEDIAN} \left( \text{IF} \left( \text{SAFFL} & \text{ARMCD}=\text{"Y"} \& \text{UPPER} \left( \text{G$2} \right) \right) \right), \text{"0"} \right) \\
C12 &= \text{TEXT} \left( \text{MIN} \left( \text{IF} \left( \text{SAFFL} & \text{ARMCD}=\text{"Y"} \& \text{UPPER} \left( \text{G$2} \right) \right) \right), \text{"0"} \right) \\
C13 &= \text{TEXT} \left( \text{MAX} \left( \text{IF} \left( \text{SAFFL} & \text{ARMCD}=\text{"Y"} \& \text{UPPER} \left( \text{G$2} \right) \right) \right), \text{"0"} \right)
\end{align*}
\]

Subsetting has already been described above but here two more functions are being here; first TEXT which takes a number and allows for a “format” to be applied, e.g. “0” rounds to an integer, and CONCAT which will concatenate text strings together. Note that for Mean the function AVERAGE is used, and Standard Deviation uses the STDEV.S function (this provides the Sample Standard Deviation).

For cells D9 to D13 and E9 to E13 following changes were made to the formulas above:

\[
\begin{align*}
D9-13: \text{UPPER(H$2)} \text{ instead of } \text{UPPER(G$2)} \\
E9-13: \text{SAFFL="Y"} \text{ instead of } \text{SAFFL&ARMCD="Y"&UPPER(G$2)}
\end{align*}
\]

The result of all this work is:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Age (years) [a]</td>
<td>n</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>Mean (SD)</td>
<td></td>
<td>44.0 (16.58)</td>
<td>49.5 (19.57)</td>
</tr>
<tr>
<td>11</td>
<td>Median</td>
<td></td>
<td>49.0</td>
<td>47.0</td>
</tr>
<tr>
<td>12</td>
<td>Min</td>
<td></td>
<td>23</td>
<td>31</td>
</tr>
<tr>
<td>13</td>
<td>Max</td>
<td></td>
<td>65</td>
<td>73</td>
</tr>
</tbody>
</table>

Doing the calculations outside the table area is possible but example shows that it is possible to do this directly inside the table itself.

The calculations for GENDER do use an area outside the table for counts as the following demonstrates:

\[
\begin{align*}
G15 &= \text{COUNT} \left( \text{IF} \left( \text{SAFFL} & \text{SEX} & \text{UPPER} \left( \text{ARMCD} \right) \right) \right) \text{"YM"} & \text{UPPER} \left( \text{G$2} \right), \text{SUBJID}) \\
H15 &= \text{COUNT} \left( \text{IF} \left( \text{SAFFL} & \text{SEX} & \text{UPPER} \left( \text{ARMCD} \right) \right) \right) \text{"YM"} & \text{UPPER} \left( \text{H$2} \right), \text{SUBJID}) \\
I15 &= \text{SUM} \left( G15, H15 \right) \\
G16-H16: \text{"YF"} \text{ instead of } \text{"YM"} \\
I16 &= \text{SUM} \left( G16, H16 \right)
\end{align*}
\]

Now putting this inside the table requires a second set of function calls:

\[
\begin{align*}
C16 &= \text{SUM} \left( G15, G16 \right) \\
C17 &= \text{CONCAT} \left( \text{TEXT} \left( G15, \text{"0"} \right), \text{"\,\textbf{(}}\text{\textbf{)}}\text{\textbf{)}}\right) \\
C18 &= \text{CONCAT} \left( \text{TEXT} \left( G16, \text{"0"} \right), \text{"\,\textbf{(}}\text{\textbf{)}}\text{\textbf{)}}\right) \\
C19 &= \text{TEXT} \left( G5-C16, \text{"0"} \right) \\
D16-19: \text{Reference to column H instead of G} \\
E16-19: \text{Reference to column I instead of G}
\end{align*}
\]

The end result of this is:
This is a good example of where it is easier to use some intermediate steps than to do the calculation and display in one cell.

To display the final footnote with tab name and date, the following formula is used:

\[
\text{CONCAT("Tab: ", MID(CELL("filename",A1),FIND("]",CELL("filename",A1))+1,255)," / Date: ",TEXT(TODAY(),"yyyy-mm-dd"))}
\]

By default the CELL function using the FILENAME parameter will bring the complete path and tab name but the request was for the tab name only, hence the double CELL call.

The TODAY function will get the current date.

**THE ADVERSE EVENT TABLE**

In the last section two of the three data types, continuous and categorical, were summarized – now it is the turn of the occurrence data where there may be several observations that are only counted once.

The first step is to create three new variables (flags):

1. Sort the ADAE data by SUBJID, AEBODSYS and AEDECOD

2. Create labels KSUBJID, KAEBODSYS and KAEDECOD at G1, H1 and I1 respectively

3. Set 1 to G2, H2 and I2.

4. G3:=IF(B3=B2,0,1) <- if match of SUBJID with previous record set 0, else 1

5. H3:=IF(B3&F3=B2&F2,0,1) <- if match of SUBJID/AEBODSYS with previous record set 0, else 1

6. I3:=IF(B3&F3&E3=B2&F2&E2,0,1)

7. Copy G3, H3 and I3 to rest of the rows for respective columns

8. Create Names Ranges KSUBJID, KAEBODSYS and KAEDECOD for the three new variables

Once this step is done the next step is to create the template for output, as in our data:
The Population and Table Header the same as previous.

Next the counts:

Note that rows in the intermediate area mirror what is in the actual table – this helps with calculations shorty. The Any Adverse Event counts are done by:

\[
I7:=\text{SUM(IF} (\text{AE}_\text{ARMCD}=\text{UPPER}(G$2), \text{KSUBJID}))
\]
\[
J7:=\text{SUM(IF} (\text{AE}_\text{ARMCD}=\text{UPPER}(H$2), \text{KSUBJID}))
\]
\[
K7:=\text{SUM(I7:J7)}
\]

Now bring the counts into the table and do the percentage counts:

Note that rows in the intermediate area mirror what is in the actual table – this helps with calculations shorty. The Any Adverse Event counts are done by:

\[
C8:=\text{CONCAT(TEXT}(I7, "0")", ",", \text{TEXT}(I7\times100/G$5, "0.0")", ",")")
\]
\[
D8:=E8:\text{Copy of C8}
\]

Now for the AEBODSYS counts:

\[
I9:=\text{SUM(IF} (\text{AEBODSYS} & \text{AE}_\text{ARMCD}=\text{UPPER}(G$2), \text{KAEBODSYS}))
\]
\[
J9:=\text{SUM(IF} (\text{AEBODSYS} & \text{AE}_\text{ARMCD}=\text{UPPER}(H$2), \text{KAEBODSYS}))
\]
\[
K9:=\text{SUM(I9:J9)}
\]

Counts can be brought into the main table as above.
For counts with AEDECOD, in I10 and J10, replace AEBODSYS with AEDECOD, and KAEBODSYS with KAEDECOD.

To speed things up, result cells, e.g. C10:E11, can be copied from C8.

Our final Adverse Event table is

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yum Yum Inc:</td>
<td>Protocol: NewSauceA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Table 14.3.2 -- Adverse Events By SOC and Preferred Term</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Safety Population</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Body System Preferred Term</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sauce (N=5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Placebo (N=4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total (N=9)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Adverse Event</td>
<td>5 (100.0%)</td>
<td>4 (100.0%)</td>
<td>9 (100.0%)</td>
<td></td>
</tr>
<tr>
<td>Cardiac disorders</td>
<td>1 (20.0%)</td>
<td>0 (0.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Palpitations</td>
<td>1 (20.0%)</td>
<td>0 (0.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Gastrointestinal disorders</td>
<td>5 (100.0%)</td>
<td>3 (75.0%)</td>
<td>8 (88.9%)</td>
<td></td>
</tr>
<tr>
<td>Constipation</td>
<td>2 (40.0%)</td>
<td>0 (0.0%)</td>
<td>2 (22.2%)</td>
<td></td>
</tr>
<tr>
<td>Diarrhoea</td>
<td>1 (20.0%)</td>
<td>1 (25.0%)</td>
<td>2 (22.2%)</td>
<td></td>
</tr>
<tr>
<td>Flatulence</td>
<td>1 (20.0%)</td>
<td>0 (0.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Food poisoning</td>
<td>1 (20.0%)</td>
<td>0 (0.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Haemorrhoids</td>
<td>0 (0.0%)</td>
<td>1 (25.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Nausea</td>
<td>0 (0.0%)</td>
<td>1 (25.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Vomiting</td>
<td>2 (40.0%)</td>
<td>0 (0.0%)</td>
<td>2 (22.2%)</td>
<td></td>
</tr>
<tr>
<td>General disorders and administration site conditions</td>
<td>1 (20.0%)</td>
<td>0 (0.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Fatigue</td>
<td>1 (20.0%)</td>
<td>0 (0.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Infections and infestations</td>
<td>0 (0.0%)</td>
<td>1 (25.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Muscle spasms</td>
<td>0 (0.0%)</td>
<td>1 (25.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Nervous system disorders</td>
<td>0 (0.0%)</td>
<td>1 (25.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Dizziness</td>
<td>0 (0.0%)</td>
<td>1 (25.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Headache</td>
<td>0 (0.0%)</td>
<td>1 (25.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Psychiatric disorders</td>
<td>1 (20.0%)</td>
<td>0 (0.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>1 (20.0%)</td>
<td>0 (0.0%)</td>
<td>1 (11.1%)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Percentages are based on N
Tab: T14.3.2 / Date: 2021-05-08

**VBA**

The first question that comes up when talking about VBA, or its true name Visual Basic for Applications, is what is it? In its simplest form, VBA is a programming language that is inside products like Microsoft Word that allow for the building of user defined functions and automating processes, closely related to the programming language Visual Basic and its predecessors, Quick Basic and WordBasic. VBA programs can be fixed to a menu button, a keyboard shortcut, or a window icon outside the Microsoft Word environment.

Microsoft Word also has its own integrated development environment (IDE), an example of which is shown below:
It is in the IDE that VBA source can be written, edited and compiled for use in your work.

An easy way to create a VBA macro is to record keystrokes that it can stored inside the spreadsheet. This is a way of programming for repeating analysis and can be edited in such a way that allowance for new data is possible.

A tutorial on VBA programming is beyond the scope of this paper.

**ADVERSE EVENT TIMELINE PLOT**

This is a very specialized plot that can display a set of Adverse Events over time, typically by patient. The type of plot that is needed is almost like a Gantt Chart but can be more easily thought of as a stacked horizontal bar chart.

Let's look at some data:

<table>
<thead>
<tr>
<th></th>
<th>BJID</th>
<th>ARMCD</th>
<th>ASEQ</th>
<th>AEDECOD</th>
<th>AEDURL</th>
<th>AEENDY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>101</td>
<td>SAUCE</td>
<td>1</td>
<td>Flatulence</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>101</td>
<td>SAUCE</td>
<td>2</td>
<td>Constipation</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>101</td>
<td>SAUCE</td>
<td>3</td>
<td>Fatigue</td>
<td>7</td>
<td>21</td>
</tr>
</tbody>
</table>

It is most important to calculate AEDUR (AE Duration) as AEENDY (AE End Day) – AESTDY (AE Start Day) which may mean a duration of 0 if the subject had the start and stop days the same.

To produce the plot for the three Adverse Events above:

1. Highlight the data for cells D2:F4
2. Select a stacked horizontal bar chart
3. Set the first series to ‘No Fill’ and ‘No Border’
4. Change the horizontal line to a start day of 1 or greater

An example output using this method for the data selected above is shown below:

Full customization using Excel options is possible – find a customization that is useful for your display and save it. Using VBA, where initially recording the keystrokes and editing the program, is something that will make an Adverse Event Timeline easy and quick to do.

CONCLUSION
In this paper it has been demonstrated that it is possible to generate basic summary output, in the examples of a Demographics and Adverse Event summary tables. However, while possible, it is does have some major challenges in that the data source is in the sheet itself and not stored in a compliant form where control is managed and secure, and there are difficulties in making the process easily expandable when updated data for a study becomes available.

CONTACT
Your comments and questions are valued and encouraged. Contact the author at:

  David Franklin
dfranklin@TheProgrammersCabin.com

Any brand and product names are trademarks of their respective companies.