Paper 073-2008

How To Generate 2, 3 and 4 Way Venn Diagrams with Drill Down Functionality within 4 minutes!

Kriss Harris, GlaxoSmithKline, UK

SUMMARY

Drill Down, Drilling Down, Graphs, HTML, Plots, Venn, Venn Diagram

ABSTRACT

When it comes to Venn Diagrams most people use a combination of Microsoft PowerPoint to generate two or three way Venn diagrams, and use filtering in Excel to count the numbers in each group. Not many people know about the four way Venn diagram and when it comes to counting the numbers in each of the 16 groups and inputting the figures into the right group it is usually done nervously and is very time consuming.

There are some websites that enable you to create 2, 3, and 4 way Venn Diagrams quite easily, however they only work with qualitative data, assume that all the elements are within the union of the groups, and do not offer you the ability to drill down in to each of the 16 groups to see the elements that are making up the groups.

This paper shows how you can combine SAS Macro, logical data steps, DSGI, and HTML to generate 2, 3 and 4 Way Venn Diagrams with Drill Down Functionality very quickly. This algorithm is very useful when looking at results of transcriptomic experiments because of the huge volume of data.

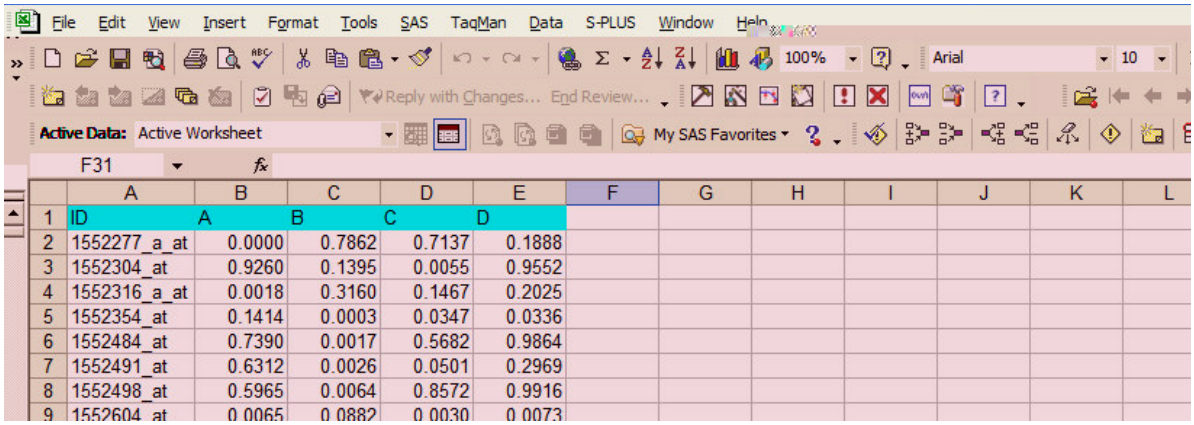
INTRODUCTION

Venn diagrams where introduced in 1883 by John Venn (1834-1923) the Hull born philosopher and mathematician. They are a great wait to visualise elements that are unique to only one group and elements that intersect with other groups, and they are symmetrical. The number of groups in an order *n* Venn diagram = 2^n (including the group outside the diagram,).

THE DATA

The data that was used for the Venn diagram SAS Macro has five columns of values and 4229 rows, one row for each probe set used. The columns A, B, C, D are (Dunnet's adjusted) p-values of four treatments compared back to control. The Macro assumes that the data is numeric, and has five column headers named **ID**, **A**, **B**, **C** and **D**.

Figure 1 The data



	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	A	B	C	D							
2	1552277_a_at	0.0000	0.7862	0.7137	0.1888							
3	1552304_at	0.9260	0.1395	0.0055	0.9552							
4	1552316_a_at	0.0018	0.3160	0.1467	0.2025							
5	1552354_at	0.1414	0.0003	0.0347	0.0336							
6	1552484_at	0.7390	0.0017	0.5682	0.9864							
7	1552491_at	0.6312	0.0026	0.0501	0.2969							
8	1552498_at	0.5965	0.0064	0.8572	0.9916							
9	1552604_at	0.0065	0.0882	0.0030	0.0073							

MACRO-VARIABLES

After importing the data the Macro uses %LET statements to create macro-variables instructing the Venn diagram to be produced, the cut off to be used, the names of the primary key and groups, and the output directory.

Figure 2 Creating the Macro Variables

```
/* Select whether you want a 2 Way, 3 Way or 4 Way Venn Diagram */
/* EG for 2 way enter 2 */

%let venn_diagram = 4;

/* Set the P Value cut-off or any other appropriate cut off */

%let cutoff = < 0.01;


/* Set the group names for the Venn Diagram */
/* Minimum requirements for a Venn Diagram
are to fill in variables id, GroupA and GroupB */

%let id = Affy_ID;
%let GroupA = TreatmentA;
%let GroupB = TreatmentB;
%let GroupC = TreatmentC;
%let GroupD = TreatmentD;

/* Enter in the location you want all the output files to go */
%let out_location = C:\Venn_Diagrams;
```

IDENTIFYING THE ELEMENTS FOR EACH GROUP OF THE FOUR WAY VENN DIAGRAM

As %let venn_diagram = 4, the macro-variable was instructed to produce the 4 Way Venn diagram. Therefore the program will follow the %DO statement commands for producing the 4 Way Venn diagram and start off by using functions and logical operators to identify the elements (in this case probe sets) that belong in the 15 groups within the diagram (the union of Group A, B, C and D).

Figure 3 Identifying the elements for each group

```
/************************** 4 Way Venn Diagram ********************************/


%IF &venn_diagram = 4 %THEN
%DO;

/* Counting the overlap */

data data_reformatted2;
set data_reformatted;
if A ne . and B ne . and C ne . and D ne . then do;
            if A  &cutoff and B  &cutoff and C  &cutoff and D  &cutoff then ABCD = 1;
            else ABCD = 0;
end;

if A ne . and B ne . and C ne . then do;
            if A  &cutoff and B  &cutoff and C  &cutoff and ABCD ne 1 then ABC = 1;
            else ABC = 0;
end;
```

COUNTING THE ELEMENTS IN EACH GROUP

After the Macro identifies the elements in each group it uses PROC UNIVARIATE to sum up the number of elements in each group.

Figure 4 Counting the elements in each group

```
proc univariate data = Data_reformatted2;
var ABCD ABC ABD ACD BCD AB AC AD BC BD CD A1 B1 C1 D1;
output out = data_sum sum = sum_ABCD sum = sum_ABC sum = sum_ABD sum = sum_ACD sum = sum_BCD
sum = sum_AB sum = sum_AC sum = sum_AD sum = sum_BC sum = sum_BD sum = sum_CD sum = sum_A1
sum = sum_B1 sum = sum_C1 sum = sum_D1;
run;
```

The total number of element within the diagram i.e. the union of Groups A, B, C, and D, and the total number of elements in the dataset i.e. the universal set are then calculated. This is used to identify the number of elements that fall outside the union.

COUNTING THE ELEMENTS THAT FALL OUTSIDE OF THE UNION

Using the fetch function the values of the total number of elements within the union and the universal set are fetched from the appropriate datasets and assigned to a macro-variable. The total number of elements that fall outside the diagram is then calculated by using %eval to evaluate the arithmetic expression of the number of elements in the universal set – the number of elements within the union.

Figure 5 Counting the elements that fall outside the union

```
/* Calculating the total number of unique ids - so that I can calculate
the number that falls outside of the groups*/
%let dsid=%sysfunc(open(id_count,i));
                        %let num_TN=%sysfunc(varnum(&dsid,count_id));
                        %let rc=%sysfunc(fetch(&dsid,1));
                        %let TN=%sysfunc(getvarn
                (&dsid,&num_TN));
                        %let rc  = %sysfunc(close(&dsid));

/* Calculating the total number of values that fall within the groups */
%let dsid=%sysfunc(open(data_sum2,i));
                        %let num_TI=%sysfunc(varnum(&dsid,totalinside));
                        %let rc=%sysfunc(fetch(&dsid,1));
                        %let TI=%sysfunc(getvarn
                (&dsid,&num_TI));
                        %let rc  = %sysfunc(close(&dsid));

/* Calculating the total numbers that fall outside all of the groups */

%let TO = %eval(&TN - &TI);
```

FETCHING THE SUMS FROM THE DATASETS

Again using the fetch function the total number of elements within each group will be fetched so that they can be displayed on the Venn diagram.

Figure 6 Fetching the sums from the datasets

```
/* Assinging the sums to macro variables */

%let dsid=%sysfunc(open(data_sum2,i));
                %let num_A1=%sysfunc(varnum(&dsid,sum_A1));
                %let rc=%sysfunc(fetch(&dsid,1));
                %let A=%sysfunc(getvarn
        (&dsid,&num_A1));
                %let rc  = %sysfunc(close(&dsid));

%let dsid=%sysfunc(open(data_sum2,i));
                %let num_B1=%sysfunc(varnum(&dsid,sum_B1));
                %let rc=%sysfunc(fetch(&dsid,1));
                %let B=%sysfunc(getvarn
        (&dsid,&num_B1));
                %let rc  = %sysfunc(close(&dsid));
```

SPECIFYING THE OUTPUT FILE AND THE GRAPH PROPERTIES

Using FILENAME outfile the output location of the Venn Diagram is specified as *4 Way Venn Diagram.gif* (within the original output location). The GOPTIONS procedure was used to set the properties for the Venn diagram including specifying the file type and length and width of the diagram.

Figure 7 Specifying the output file and graph properties

```
/* Outputting the Venn Diagram - Change the output file location */
ods html;

filename outfile "&out_location\4 Way Venn Diagram.gif";

goptions RESET=ALL device = emf gsfname = outfile FTEXT = "Arial" gsfmode=replace
colors=(black, red, green, yellow, blue, magneta, cyan, orange)
hsize = 8 in vsize = 6 in border
XPIXELS=800 ypixels = 600 ;
```

DRAWING THE VENN DIAGRAMS

Data Step Graphics Interface (DSGI) was used to draw the Venn diagrams. The GSET function was used to set the Line Colour and the Line Width of the Venn Diagram Ellipse. For example the first ellipse is Hollow with a Red Outline. The GDRAW function is used to draw the ellipse. The syntax for the hollow ellipse is: **return-code-variable** = **GDRAW('ELLARC', x, y, major, minor, start, end, angle);**

The argument definitions of the first ellipse are that is has an origin of (80, 50 (+ offset)) with a length of 160 and a width of 60, it is 360 degrees and has an angle of 50 degrees.

Figure 8 Drawing the Venn diagram

```
/* Drawing the Ellipse Plots with the Counts */


%let offset = -8;
DATA _NULL_;
/*Initialize DSGI*/
RC = GINIT ();
/*Open a graphic segment*/
RC = GRAPH ('CLEAR',  'GDSENLAR');


/* First Ellipse */

/*No. 2 color, Red*/
RC = GSET ('LINCOLOR', 2);
RC = GSET ('LINWIDTH', 2);


   /*Draw a Hollow Red Ellipse, from 0 to 360 degree at coordinate (80, 50) with
length 160 and width 60 at 50 degrees*/
RC = GDRAW ('ELLIPSE', 80, 50 &offset, 80, 30, 0, 360, 50);
```

DISPLAYING THE TITLES AND THE SUMS ON THE VENN DIAGRAM

The GSET function was used to set the font size, text and colour, and the GDRAW function was used to draw the text that was stored in the macro-variables at the appropriate origins.

Figure 9 Displaying the titles and the sums of the Venn diagram

```
/* Text */

RC = GSET ('TEXHEIGHT', 3);

/*Write down text, “&GroupA” starting from coordinate, (28, 60)
*/

RC =GSET('TEXCOLOR', 3);
RC = GDRAW ('TEXT', 0, 37 &offset, "&GroupA");

RC =GSET('TEXCOLOR', 1);

RC = GDRAW ('TEXT', 28, 60 &offset, "&A");

RC = GDRAW ('TEXT', 43, 90 &offset, "&B");

RC = GDRAW ('TEXT', 69, 90 &offset, "&C");

RC = GDRAW ('TEXT', 88, 60 &offset, "&D");
```

CREATING THE AREAS FOR DRILLING DOWN

Put statements were used to create HTML image maps that specify areas around the numbers in each group of the Venn diagram. When the area is selected it is linked to the elements that make up the group.

Elaborating on the method it shows that in the **BODY** of the *4 Way Venn Diagram.html* file the **IMG** tag is used to reference the *4 Way Venn Diagram.gif* file and to define which image map to use, in this case *fourvenn*. The **MAP** tag defines the image map, and inside the **AREA** tag specifies the coordinates of the clickable area and the file to link it to once the area is selected.

Figure 10 Creating the drilling down areas

```
/* Drilling down */

/* Specifying the regions */
filename four "&out_location\4 Way Venn Diagram.html";

/* Have four put quotes around the quotes so that the outfile location is specified */

data _null_;
  file four;
  put '<HTML>';
  put '<BODY>';
  put '<IMG SRC="""&out_location\4 Way Venn Diagram.gif""" usemap = "#fourvenn" />' ;
  put '<map id ="fourvenn" name="fourvenn">';
  put '<area shape ="rect" coords ="155,279,193,255" href = """&out_location\4 Way Tables.html#anchor"""/>';
  put '<area shape ="rect" coords ="241,108,280,83" href = """&out_location\4 Way Tables.html#anchor1"""/>';
  put '<area shape ="rect" coords ="381,108,426,83" href = """&out_location\4 Way Tables.html#anchor2"""/>';
  put '<area shape ="rect" coords ="492,279,535,255" href = """&out_location\4 Way Tables.html#anchor3"""/>';
```

CREATING THE DRILL DOWN LISTS

PROC SQL enclosed in a Macro was used to select the elements that make up a group for each group within the union, using the parameter *list*. PROC SQL was then used to list the elements that fall outside the group.

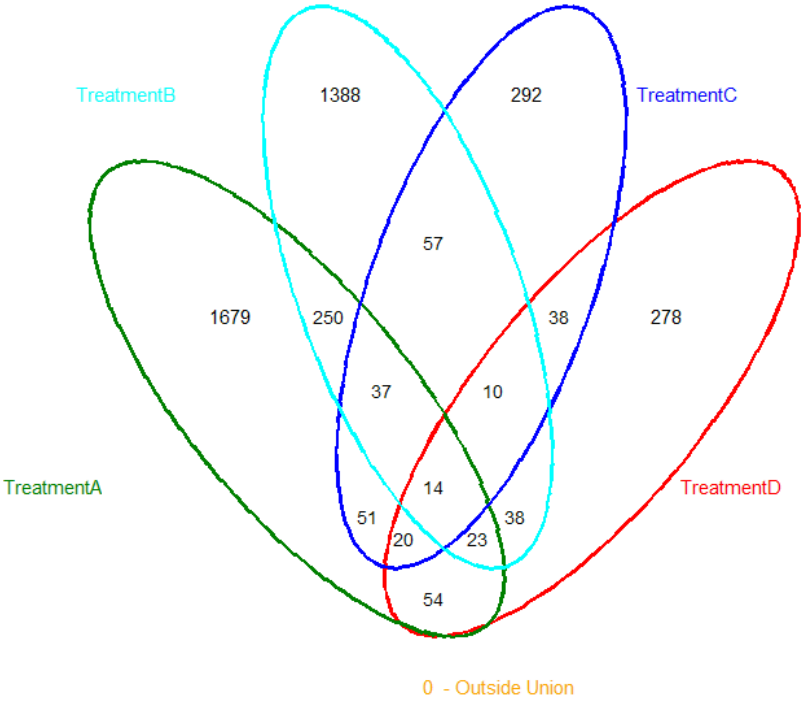
Figure 11 Creating the drill down lists

```
%macro table(list, title);

title1 "&title.";
proc sql;
select id
from Data_reformatted2
where &list. = 1
order by id;
quit;

%mend;
%table(A1, &GroupA);
%table(B1, &GroupB);
%table(C1, &GroupC);
%table(D1, &GroupD);
%table(AB, &GroupA and &GroupB );
%table(AC, &GroupA and &GroupC );
%table(AD, &GroupA and &GroupD );
%table(BC, &GroupB and &GroupC );
%table(BD, &GroupB and &GroupD );
%table(CD, &GroupC and &GroupD );
%table(ABC, &GroupA and &GroupB and &GroupC );
%table(ABD, &GroupA and &GroupB and &GroupD );
%table(ACD, &GroupA and &GroupC and &GroupD );
%table(BCD, &GroupB and &GroupC and &GroupD );
%table(ABCD, &GroupA and &GroupB and &GroupC and &GroupD );

title1 "Others";
proc sql;
select id
from Data_reformatted2
where A1 = B1 = C1 = D1 = AB = AC = AD = BC = BD = CD =
ABC = ABD = ACD = BCD = ABCD = 0 ;
quit;
```

FINAL VENN DIAGRAM

The final 4 Way Venn Diagram shows the relationship between the four groups.

Figure 12 4 Way Venn Diagram



DRILL DOWN EXAMPLE

In the HTML document if you selected any of the regions you will be linked to the a file that shows the elements that make up that group. This is not the HTML document, but for example if you selected the region that intersects Treatment B, C and D i.e. where the number is 10 you will be linked to:

Figure 13 Drill down example

TreatmentB and TreatmentC and TreatmentD

ID
1564248_at
1570491_at
205109_s_at
205707_at
208187_s_at
222550_at
227091_at
233483_at
243302_at
243990_at

CONCLUSION

Venn diagrams are very useful for visualising the relationships between groups. This paper has demonstrated how the useful technique with drill down functionality can be implemented in minutes using SAS.

The columns in the data are assumed to be numerical if someone has the time and passion they can edit the macro so it accepts qualitative data!

ACKNOWLEDGEMENTS

I want to thank Adam Taylor for prompting me to start the Venn diagram SAS Macro by asking me if it was possible to do Venn diagram calculations in SAS and I would like to thank Mark Lennon for showing interest in plotting the Venn diagrams in SAS.

REFERENCES

3 Way Venn Diagram - http://ftp.sas.com/techsup/download/sample/graph/other-venn.html
HTML Image Maps - http://www.quackit.com/html/tutorial/html_image_maps.cfm
Making Presentations More Fun with DATA Step Graphics Interface (DSGI) - http://www2.sas.com/proceedings/sugi31/056-31.pdf
SAS Institute
Venn Diagrams - http://en.wikipedia.org/wiki/Venn_diagram

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the author at:

Kriss Harris
GlaxoSmithKline
Mail Stop : 1B106003
Gunnels Wood Road
Stevenage SG1 2NY
Phone: 01438 76 4904
Email: Kriss.5.Harris@gsk.com