Paper 217-2009

Using SAS® to Create Proportional Venn Diagrams

Shiqun (Stan) Li, Minimax Information Services, Belle Mead, NJ

ABSTRACT

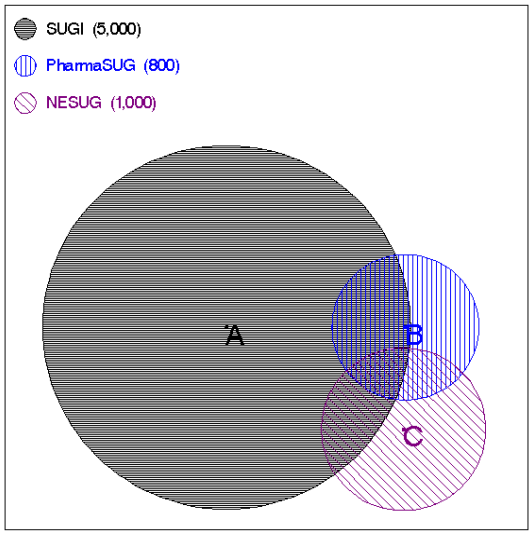
Proportional Venn diagrams have a wide application background. They can be used to visually demonstrate the magnitudes and relationships among subgroups. Within our SAS community, there has been a highly demanding for a new SAS procedure that can generate proportional Venn diagrams[1]. In this presentation, we will propose a SAS macro %pVenn. This SAS macro will be capable of drawing two-circle or three-circle Venn diagrams that are area-proportional to the size of the sets. We will also exhibit some examples of utilizing this SAS macro to generate proportional Venn diagrams for real world applications. This paper is prepared for an intermediate and advanced audience.

BACKGROUND

Venn diagrams was proposed for representing logical positions by John Venn in 1880. A proportional Venn diagrams typically has three circles representing for three subsets from the same population. Each circular area stands for the relative size of each subset (like number of count or percentage); the overlap area of two circles implies the sharing amount between the coresponding subsets. The intersection areas, or the set relationships, are particularly interested in many applications.

Before we go further, let us introduce an example of proportional Venn diagrams. Suppose we are interested in the number of attendees of our three SAS user conferences: SUGI (Set A), PharmaSUG (Set B) and NESUG (Set C). Assume, for a certain year, there were 5,000 attendees for SUGI, 800 people for PharmaSUG and 1,000 persons for NESUG. In addition, we estimate 400 people went to both SUGI and PharmaSUG, 300 joined both SUGI and NESUG, 200 attended both PharmaSUG and NESUG. That is, |A|=5,000, |B|=800, |C|=1,000, |AB|=400, |AC|=300 and |BC|=200. (Please note the numbers here are not to be accurate. They are for the example purpose only.)

To represent these three conference attendees with the assumed characteristics, a proportional Venn diagram will be a nice choice. It can be visually displayed as below:



In this diagram, circular area of A, representing |A| which is the number of SUGI attendees, is the largest, much larger than |B| and |C|. The intersected area of A and B in the graph is proportional to the size of AB. That is, |AB| (those who attended both SUGI and PharmaSUG) is proportionally represented by the joined area of A and B. Similarly, the intersection areas of A and C is proportional to the number of attendees to both SUGI and NESUG; and the overlay area of B and C stands for the number of people who went to both PharmaSUG and NESUG. From this diagram above, it is easy to demonstrate the relative sizes and the overlay relationships of these three conferences.

This kind of Venn diagrams can be an excellent tool in epidemiology, healthcare and marketing. In epidemiology, if we let the 3 subsets A, B and C represent 3 kinds of diseases respectively, the graph will present the relationships of the 3 diseases. We will demonstrate some examples on this later.

PROPERTIES

Venn diagram of circular shape is a very interesting topic. Proportional Venn diagrams and the properties were well discussed in several previous studies[2, 3, 4]. The discussion on the properites will be omited here.

The properties of Venn diagram of other shapes (e.g. rectangle) had also be addressed in several studies. But in this presentation, we will focus our interest on Venn diagrams of circular shape.

ALGORITHM

We here present an algorithm to create an area-proportional Venn diagram of 3 sets: A_1, A_2 and A_3. First, assume $|A_1|= a_1$, $|A_2|= a_2$, $|A_3|= a_3$, $|A_1A_2|= a_{12}$, $|A_1A_3|= a_{13}$, $|A_2A_3|=a_{23}$ and $|A_1A_2A_3|=a_{123}$. Without loss the generality, we further assume A_1, A_2 and A_3 are subsets of a population P, and with $0<=a_i, a_{ij}, a_{ijk}<1$ (where$1<=i<j<k<=3$). Based on these assumptions, we propose the following SAS algorithm to draw a proportional Venn diagram:

a) Calculate the radius of the circle for each set: $r_i=\text{sqrt}[a_i/(2\pi)]$, where $|A_i|=a_i$.
b) Draw the first circle A_1 with center (0, 0) and radius r_1.
c) Compute the distance d_{12} for A_1 and A_2, given r_1, r_2 and $a_{12}=|A_1A_2|$.
d) Draw the second circle A_2 with center (d_{12}, 0) and radius r_2.
e) Estimate the distance d_{13} for A_1 and A_3, and distance d_{23} for A_2 and A_3.
f) Find the center (x_3, y_3) of the third circle A_3. With some basic geometry knowledge, it is easily to obtain that the center of the third circle A_3 can be expressed in the algebraic fomula as $x_3=(d_{12}^2+ d_{13}^2- d_{23}^2)/(2d_{12})$ and $y_3=\text{sqrt}(d_{13}^2-x_3^2)$.
g) Draw the third circle A_3 with center (x_3, y_3) and radius r_3.
h) Add the legend and others as required to the diagram.

SAS MACRO %pVenn

Based on the algorithm above, we programmed a SAS macro %pVenn that can be used to generate proportional Venn diagrams of 1 circle, 2 circles or 3 circles. The structure of macro %pVenn looks like:

```
%macro pVenn3(DSN=, Var1=, Var2=, Var3=,
              Name1=, Name2=, Name3=, Color1=, Color2=, Color3=,
              Base=0, Scale=0.80, Fmt=percent8.1);
```

where the input variables are:

DSN: the input SAS dataset

Var1, Var2 and Var3: the variable names for the three sets

Name1, Name2 and Name3: the names or labels for set A, B and C. If Names are not specified, legend will not display.

Color1, Color2, Color3: the colors for the three subsets, default as Black, Blue and Red.

Base: When Base=0, the square around the Venn diagram will be just a frame and has no statistical meaning. But when Base>0, the frame or the square around the Venn diagram will be served as the size of population P that Set A, B and C are based upon. For example when Base=1, a square frame will be drawn to represent |P|=100%. By default, Base=0.

Scale: SCALE can be used to adjust the relative sizes of the 3 circles, when BASE=0. But if Base=1 is specified, SCALE is void.

FMT: This is the format used to display the statistics in the legend. If FMT=<*blank*>, the data values will not show in the legend (another way to suppress the legend).

This macro does not require very fancy SAS functionalities. We utilize ANNONATE and PROC GSLIDE to create the circles. ANNONATE is indeed a amazing functionality in SAS/Graph and has provided a lot of flexibilities in SAS/Graph. Of course, besides the SAS techniques, our algorithm needs some numerical computation skills as well. It involves solving nonlinear equations in order to find the distance between two circles[2, 3, 4]. We also have a version of S-Plus® function that can generate proportional Venn diagrams. For more details about our SAS macro or our S-plus function, please refer to the Appendix or contact the author.

There is a SAS sample code that can draw non-proportional Venn diagrams [7]. This code can only generate Venn diagrams of 3 circles, but the areas of the circles and the intersections are not proportional to the size of subsets and their intersected sets. However, this sample code was a good start for our %pVenn macro.

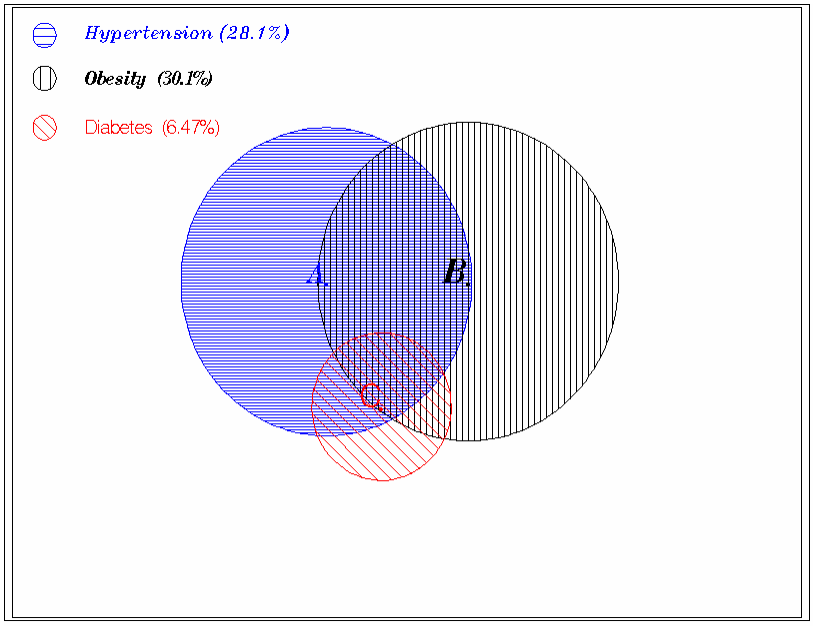
APPLICATION EXAMPLES

In order to demonstrate the functions of macro %pVenn, we will show some simple applications in this section. In the following example graphs, the data mostly comes from NHANES database, a National Health and Nutrition Examination Survey database which can be found from this link: http://www.cdc.gov/nchs/nhanes.htm. Please note the numbers in the graphs are intended for demonstration purposes only.

First, a typical Venn diagram example for three subsets intersected with each other. With our macro %pVenn, the syntax will be like:

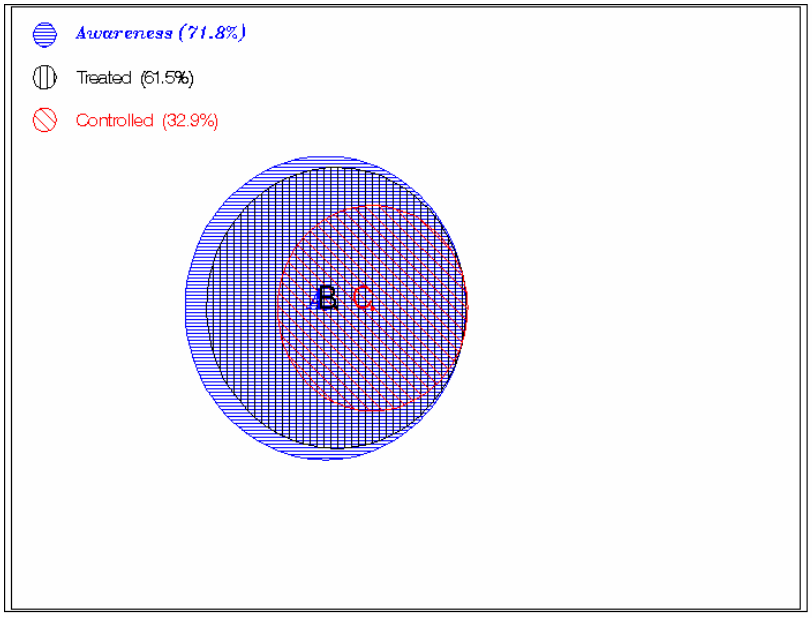
```
%pVenn3(DSN=NHANES, Var1=HTN, Var2=Obesity, Var3=Diabetes,
        Name1=%str(Hypertension), Name2=%str(Obesity), Name3=%str(Diabetes),
        Fmt=percent8.1, base=0, scale=.8);
```

The output of the proportional Venn diagram for the three variables: Hypertension, Obesity and Diabetes, will be displayed as:

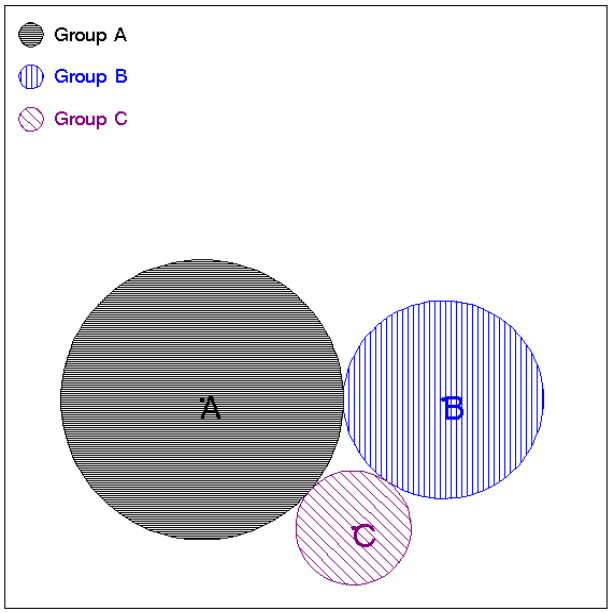


The relationships of the given three sets can have many cases. The 3 sets can be intersected like the example above; they can be totally overlaid with one set containing another set; they also can be three disjoined sets. Our macro is able to produce the graphs for all those situations. Below is an example of 3 sets with A contains B and B contains C.

```
%pVenn3(DSN=NHANES, Var1=Awareness, Var2=Treated, Var3=Controlled,
      Name1=%str(Awareness), Name2=%str(Treated), Name3=%str(Controlled),
      Fmt=percent8.1, base=0, scale=.8);
```

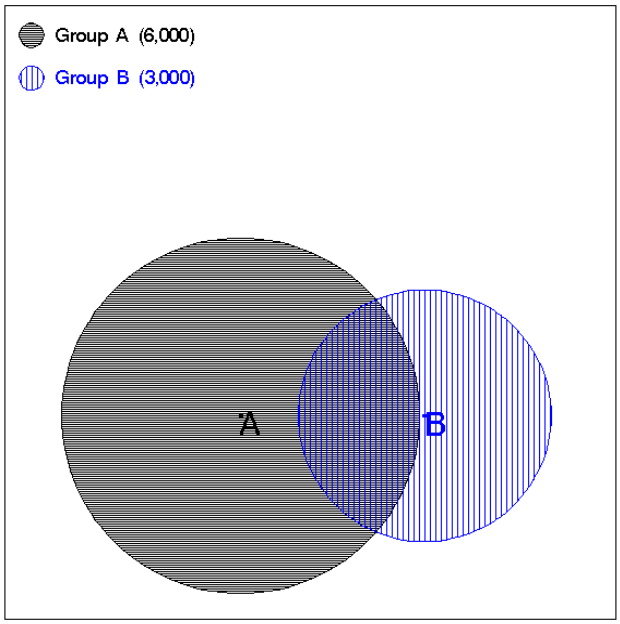


The next diagram comes from an example of three sets without intersection, or disjoined. An example is: A=Male, B=Female and C=Unknown. This kind of graph has not much interest in the real world, since there are not relationships among the three subsets. We just want to point out that %pVenn can generate this type of diagrams as well.



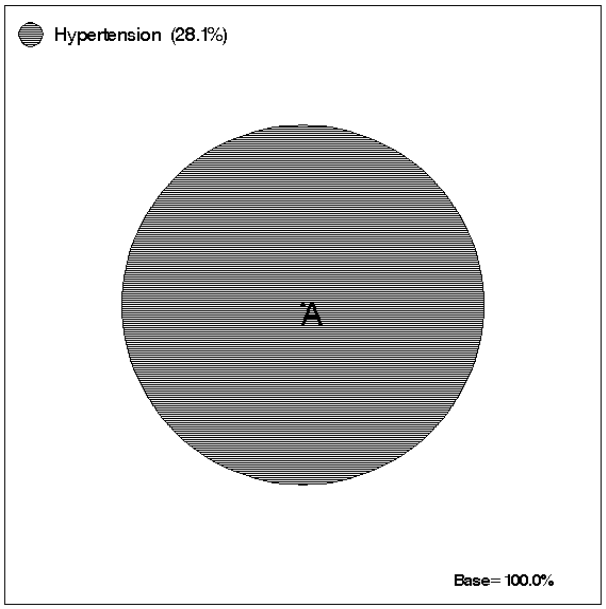
Sometimes, we may be only interested in the relationships between two subsets. Without any modification, macro %pVenn can be used to generate a two-set Venn diagrams. The syntax of using this macro for this situation is (just omit the third variable name):

```
%pVenn3(DSN=Example, Var1=G1, Var2=G2, Var3=, Name1=%str(Group A),
Name2=%str(Group B), Name3=%str(), Fmt=8., base=0, scale=.8);
```



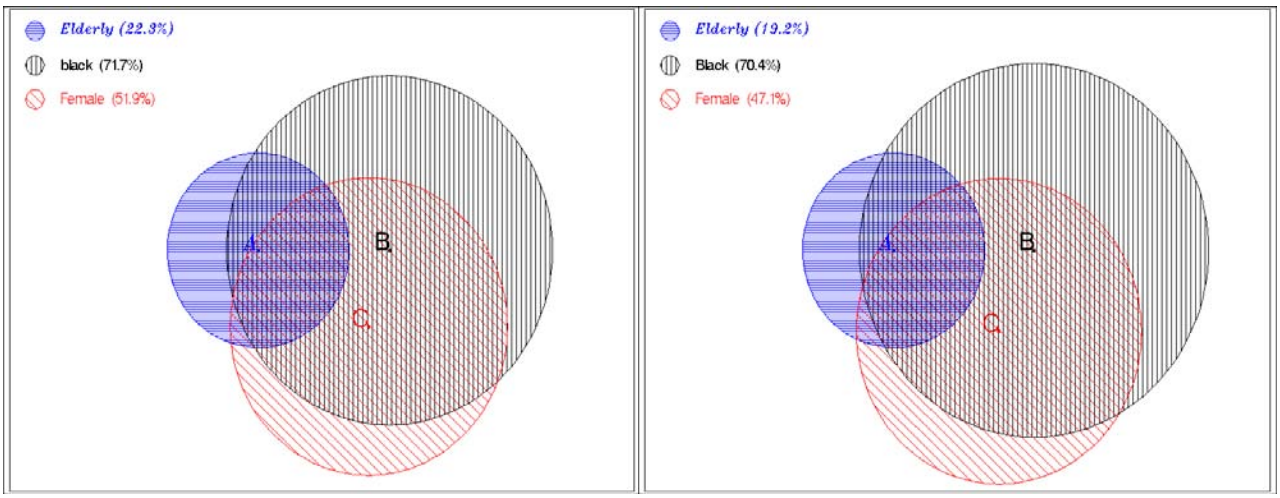
Similarly, %pVenn can produce a Venn diagram with only one set. Please note, in the sample graph below, there is a background square with area=100%. The macro input for making the diagram below is like:

```
%macro pVenn3(DSN=NHANES, Var1=HTN, Var2=, Var3=,
          Name1=%str(Hypertension), Name2=, Name3=, Color1=black, Color2=,
          Color3=, Base=1, Scale=1, Fmt=percent8.1);
```



As we can see from the examples above, Venn diagrams are excellent techniques to visually present the relationships among subsets. They can be good methods for comparing the characteristics among groups also. Below is a comparison example of three baseline characteristics between two treatment arms.

Treatment A **Treatment B**



LIMITATIONS AND DISCUSSIONS

The intersected area of all 3 circles is called triangle circular. Given 3 sets A, B and C, the values |A|, |B|, |C|, |AB|, |AC| and |BC| will be enough to determine the corresponding locations of the 3 circles representing the 3 sets. The intersection area |ABC| is redundant. Therefore, it must be noted the overlay triangle circular may not be exactly proportional to the intersection of all three sets A, B and C. This is a limited nature for Venn diagrams of circle shape. There should have a footnote to the graph to notify the readers if there is a significant difference in this triangle circular.

It is also necessary to point out there are some certain situations that cannot be represented by 3-ring proportional Venn diagrams. A case like this is: Subset A: Male=50%, Subset B: Female=50%, Subset C: certain disease with prevalence=25% (15% for Male and 10% for Female). For this circumstance, it is impossible to use a Venn diagram with 3 circles to display the relationships of the data. This is very obvious in geometry. When this type of situations are encountered, macro %pVenn will issue a warning message and stop further procedures. This problem could be avoided by choosing a different geometry shape like rectangle for the Venn diagram. A Venn diagram of other shape will be integrated into our macro later.

To this point, we only discuss Venn diagram with up to 3 rings. If there are more than 3 sets to be displayed, a Venn diagram of circle shape may not be a good choice [3]. For most situations, it is impossible to position the location for the 4^{th} circle for the 4^{th} set. Other approaches with alternative shape will be required for cases with more than 3 sets. But fortunately, a 2-ring or 3-ring Venn diagram can serve most of our interests. If there is a need to demonstrate 4 sets at the same time, we could produce multiple 3-ring diagrams and display them in the same page.

CONCLUSION

Proportional Venn diagrams are powerful tools to display the relationships of subsets. We have proposed an algorithm and provided a SAS macro %pVenn that can create 2-ring or 3-ring Venn diagrams. We have utilized this macro to generate graphs and reports on various projects. It is indeed a nice technique to present visual-pleasing reports. We have also developed a S-Plus function pVenn{} that can create proportional Venn diagrams in S-Plus. Interested SAS users or S-Plus users are welcome to contact us for more information and for the codes.

CONTACT INFORMATION

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

Shiqun (Stan) Li
Minimax Information Services
(908)240-8229
shiqun@gmail.com

TRADEMARKS:

REFERENCES

1. SAS Surveys/Ballots: On “Provide a Procedure to Generate a Venn Diagram”:
http://support.sas.com/techsup/feedback/sasware_ballot01/ballot01.survey.htm
http://www2.sas.com/proceedings/sugi26/front/2001_SASwareBallot.pdf
http://ftp.sas.com/techsup/download/sasware/02results.html
http://ftp.sas.com/techsup/download/sasware/ballot2003.pdf

2. Chow, S. and Ruskey, F., *Drawing Area-Proportional Venn and Euler Diagrams*, 11th International Symposium on Graph Drawing, Perugia, Italy, Lecture Notes in Computer Science, 2912 (2003) 466-477.

3. Stirling Chow and Peter Rodgers: Extended Abstract: Constructing Area-Proportional Venn and Euler Diagrams with Three Circles, http://www.cs.kent.ac.uk/pubs/2005/2354/content.pdf

4. Fewell, M. P., Area of Common Overlap of Three Circles. http://nla.gov.au/anbd.bib-an000041411907

5. J. Stoer and R. Bulirsch, Introduction to Numerical Analysis. Springer 2002

6. Alfio Quarteroni, Riccardo Sacco and Fausto Saleri, Numerical Mathematics. Springer 2006

7. SAMPLE: Generating the Venn Diagram, ftp.sas.com/techsup/download/sample/graph/other-venn.html

APPENDIX

It would be clumsy and lengthy to attach the whole macro %pVenn here. Instead, we will just append the core codes here.

1. The nonlinear equation of the 4 variables: radii r_1, r_2, the distance d and the intersected area A between the two circles:

```
%macro func(r1, r2, d, A);
  &r1*&r1*arcos((&d*&d+&r1*&r1-&r2*&r2)/(2*&r1*&d)) +
  &r2*&r2*arcos((&d*&d+&r2*&r2-&r1*&r1)/
  (2*&d*&r2))-0.5*sqrt((&r1+&r2-&d)*(&r1+&d-&r2)*(&r2+&d-&r1)*(&r1+&r2+&d))-
&A;
%mend func;
```

2. Macro to find the distance between two circles, given radii r_1, r_2 and the intersected area A. We use the bisection algorithm to solve for the distance d from the non-linear equation above. Other methods like Newton iteration can be employed for this purpose also.

```
%macro bisection(r1,r2,A, root ) ;
T=1E-6;
_a=abs(&r1-&r2)+T; _b=(&r1+&r2)-T;
DO while(1);
      _c=(_a+_b)*0.5;
      F_a=%func(&r1,&r2,_a, &A.);
      F_c=%func(&r1,&r2,_c, &A.);
      if abs(F_a) le T then do; &root.=_a; leave; end;
      if abs(F_c) le T then do; &root.=_c; leave; end;
      if F_a*F_c lt 0 then _b=_c;
      else _a=_c;
      if _b-_a<T then do;
        &root.=_c; leave;
      end;
END;
%mend bisection;
```

3. After the centers of the 3 circles are located, we use this macro to generate an annotate data set.

```
*** To draw pie, center, Label and legend;
%macro Pies(x=,y=,size=,color=,PieStyle=, TFont=, labels=, Letter=,Lx=,Ly=);
```

```
/* (x,y) is the center, size=radius,
   Letter: a letter to label the circle
   PieStyle=PnNa: n=line density, a=angle of lines/patterns
   For legend:
   TFont=font for labels, labels=text label for the legend
   (Lx, Ly): center for the legend circle
*/

function='pie'; x=&x.; y=&y.; size=&size; color="&Color."; line=0;
style="&PieStyle.";rotate=360;output; *** Circle with pattern;

function='pie'; size=.3; style='solid'; output;  *** a dot at center;

function='label'; text="&Letter."; position="A";
style="&TFont.";size=5; output;  *** labelling the circle;

*** legend: a small circle and a text string;
function='pie'; x=&Lx.; y=&Ly.; size=2; line=0;
style="&PieStyle.";rotate=360; output;
function='label'; x=%Eval(&Lx.+4); y=%eval(&Ly.+1); text=&Labels.;
position="6";
style="&TFont.";size=3; output;
%mend Pies;
```

4. With the annotate dataset, it is easy to create the Venn diagram now:

```
title h=2 "&title.";
proc gslide annotate=AnnoSet frame;
footnote1 h=1 "&footnote.";;
run;
quit;
```