venndiag: Drawing Venn Diagrams

Jens M.Lauritsen, Initiative for Accident Prevention, County of Fyn, Denmark, jml@dadlnet.dk

venndiag produces a so-called Venn diagram based on variables in a dataset

The Venn diagram consists of a number of rectangles each corresponding to one of the variables in varlist. The rectangles are arranged such that they overlap and delimit areas. In each area the counts of records is shown for the relevant combination of varlist.

With two variables A and B, the counts of records (A== 1 & B== 1) is placed in the overlapping area of A and B. In area A (non A+B) the count (A== 1 & B!= 1) is placed and in area B (non A+B) the count (A!=1 & B==1).

The command has three types of output:
One which creates the combinations of the variables and presents the counts of this in the log file.
One which contains the actual diagram presented on the **left side** of a graph window
One which presents labelling and contents information on the **right side** of a graph window

Venndiag could be used, for example, to
- show number of persons having different symptoms indicated in the three variables. E.g astma, hayfever and eczema
- show in a household survey the numbers of having cats, dogs and birds
- count specific diagnoses placed in one or several variables
- combine variables and achieve frequency counts in logfile and in a new variable
- create a new combined variable of 2-4 variables (with or without missing)

Figures are produced if you run the do file: ***venntest.do***, which produces simulated data and the figures for inclusion here. If you run the ***venndiag*** with the option **print** graphs will be printed directly on win-95 operating systems.

Syntax

venndiag varlist [if exp] [in range] [,label() show() missing gen() list() print saving() c1() c2() c3() c4() noframe nograph nolabel t1itle() t2title() t3title() r1itle() r2title() r3title() r4itle() r5title() r6title()]

The varlist must contain from 2-4 numerical variables and if generating a variable, that variable must be non-existing.

Remarks

Texts in the graph are sized accordingly to the ***set textsize*** command. 100 is default. Experimentation with sizes is recommed, ***set textsize 115, set textsize 125*** etc. If set at a value above 120 some texts might be outside areas.

Disallowed variable names in dataset are: *_merge, vdx1199 and vd*_id199. If a system macro is set with the command (global S_grid = "y") a grid will be placed on the screen for further placement of texts by user.

Information is retrievable in S_* variables after execution (See start of venndiag.ado for the numbers of these)

Basis for percentages is the number of records included in the graph totally. If the user wishes percentages to be based on only those records affirmative of at least one the variables then include only records affirmative of at least one of the variables. Sometimes the user might want to show a graph with percentages based on all cases, but not showing the records which do not have at least one variable affirmative. This is accomplished with the show(x) option.

Options

show() Show in rectangles: (default: pctf) (sequence unimportant)

- **p** percent of area (area/total)
- **c** count in each area
- **t** percent of each variable (variable/total).Note: Σ variable percentages > 100 %.
- **v** variable name instead of A B C D
- **f** add footnote below explaining percentages
- **n** display counts and percentages for areas with counts=0 (not)
- **x** exclude the counts and or percentages of the non area (records in non area are still included in N)
- **all** short form for all (not **x**)

nograph Do not show graph, regardless of contents of show() and label()

label() Include only labelgroups of: (default adft) (sequence unimportant)

- **c** variable names and value counted in each variable
- **f** filename and date
- **d** overall description
- **m** indicates counts of missing in each variable
- **t** titles are shown
- **a** date of creation of graph added
- **i** Total records in file, # exluded (in/if or missing values) and number of records in graph shown
- **x** show information in titles **r1title ...r6title** (should not be used with **fdmc**)
- **all** short form for all labels (not **x**)

nolabel Disregards all labelling, regardless of contents of label

c1() **c2**() **c3**() **c4**() Specifies which value to regard as outcome in v1...v4 (default 1) must be integer

noframe Remove outer frame

missing Include all records regardless of missing variables in varlist

t1title() **t2title**() **t3title**() titles shown on graph. Default t1(*Venn Diagram*) t3(*N = #records*)

r1title() r2title() r3title() r4itle() r5title() r6title() Additional titles to shown when **x** is included in label()

saving(*filename*) Save graph to *filename* (replacing file). On Win-95/MAC also saved as wmf/pict file (see gphprint)

gen() Add variable named in () to dataset. Must be non existing

list(*variables*) List *varlis*t in call of venndiag and *variables* after combining records

print Print graph immediately (Win-95 & MAC only)

Examples

The simplest plot with no options specified will look as in figure 1. The boxes are named A, B, C (with three variables) and counts of each area placed in an appropriate place. Percentages of areas are shown in (%) and for each variable the percentage having the counted outcome is shown without parenthesis. Titles in the default mode with no options are as seen, the date of creation and datefile used plus variable labels and total N shown with the genereal title ”Venn Diagram”. Any record having a value of missing will be excluded from the graph. This will be indicated in the log file together with the counts of the possible combinations of the variables. To produce the first figure, the only option applied was *`saving(figure1)`*

```
. venndiag astma season eczema, saving(figure1)
__________________________________________________________________
____

Venn diagram of variables: astma season eczema
File: testdata.dta (Cr:16 Nov 1998 )

     Outcome    Variable and label
A:       1           astma    Astma previous year
B:       1           season    Seasonal allergic symptoms
C:       1           eczema    Current hand eczema

    4000   Records in file
      78   Records excluded by missing values
    ____
    3922   Records in Diagram:
Counts for combined variables:
------------------------------------------------------------------
----
A      |    138    4 %  (astma == 1)&(season != 1)&(eczema != 1)
B      |    100    3 %  (astma != 1)&(season == 1)&(eczema != 1)
C      |    300    8 %  (astma != 1)&(season != 1)&(eczema == 1)
AB     |    165    4 %  (astma == 1)&(season == 1)&(eczema != 1)
AC     |     11    0 %  (astma == 1)&(season != 1)&(eczema == 1)
BC     |     74    2 %  (astma != 1)&(season == 1)&(eczema == 1)
ABC    |     74    2 %  (astma == 1)&(season == 1)&(eczema == 1)
```

```
---    |   3060   78 %  (astma != 1)&(season != 1)&(eczema != 1)
-------------------------------------------------------------------
----
___________________________________________________________________
____
```

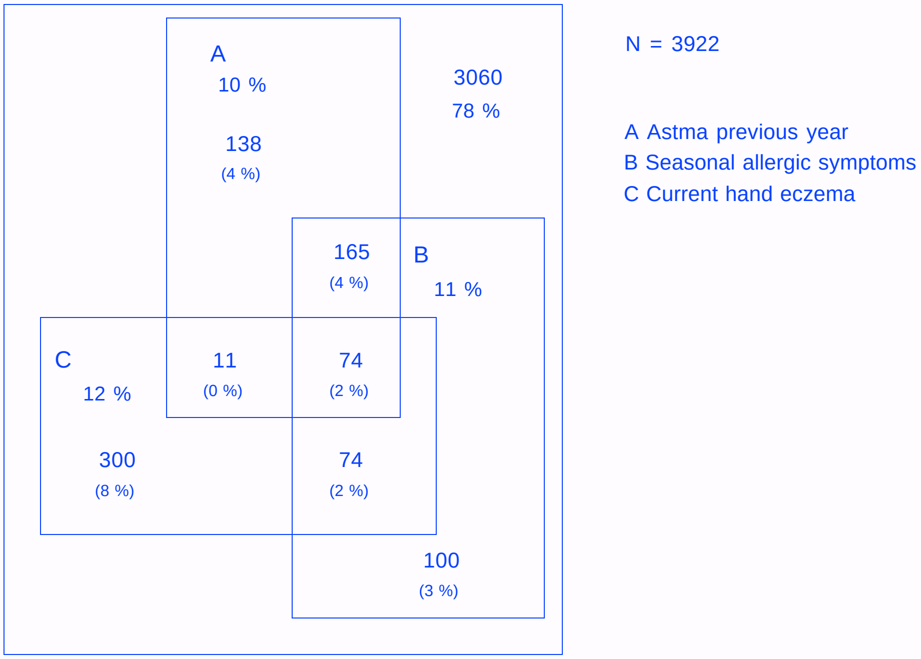
Figure 1 here

Sometimes a variable is recorded with positive answers only, i.e. a missing indicates a ”non-affirmative” answer. When this is the case, records containing missing values should be included. This is indicated by the option **missing.** In the second example a few other options are shown. Values indicated in the c1() to c4() will be used to count outcome instead of 1, and furthermore additional information on the graph will be shown with the **label(all)** and **show(all)** options. The **all** is a short form for several options, se above. With the **noframe** option, the external frame around the boxes is not shown. The log file will inform of the inclusion of missing values.

```
. venndiag eczema-atopi,  saving(figure2) noframe missing
t3(Note: 0's shown on graph) `1'/*
>    */ label(all) show(all)  t1(Venn Diagram - all information
shown on right half of graph)    /*
>    */ c1(2) c2(3) c3(4) c4(5)
___________________________________________________________________
____

Venn diagram of variables: eczema astma season atopi
File: testdata.dta (Cr:16 Nov 1998 )

... output omitted ....

4000   Records in Diagram:

     104   variables in all records contain missing values
           eczema       :     26
           astma        :     26
           season       :     26
           atopi        :     26

Counts for combined variables:
-------------------------------------------------------------------
----
A      |    274    7 %  (eczema == 2)&(astma != 3)&(season !=
4)&(atopi != 5)
... output omitted ....
```

Figure 2 here

Venn Diagram

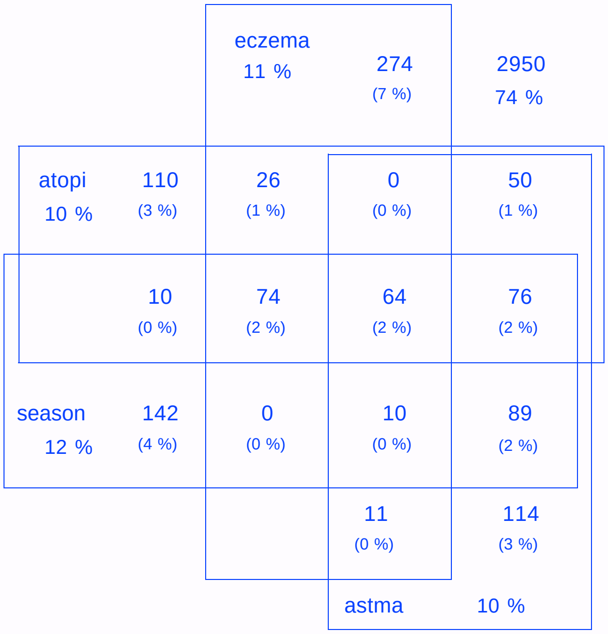


% of total (% in area of total)

25 Nov 1998

File: testdata.dta (Cr:25 Nov 1998)

Venn Diagram - all information shown on right half of graph



% of total (% in area of total)

N = 4000

Note: 0's shown on graph

A Current hand eczema
B Astma previous year
C Seasonal allergic symptoms
D Childhood atopic symtptoms

Value indicators:
A: (eczema=2)
B: (astma=3)
C: (season=4)
D: (atopi=5)

Records in file:	4000
Excluded: Miss 0 In/if: 0	0
Total Records in graph:	4000

Missing values (Records) (Included in graph)

A:(eczema=26)
B:(astma=26)
C:(season=26)
D:(atopi=26)

25 Nov 1998

File: testdata.dta (Cr:25 Nov 1998)

Defaults of labels are handy, but sometime a finetuning is desirable. If the user specifies **label(xt)** only texts put into **t1()...t3() r1() ...r6)** are added to the text. With the options in **show()** the graph contents on the left can be controlled. The bottom line description is excluded by omitting **f** from the parenthesis e.g. **show(xcpt)**

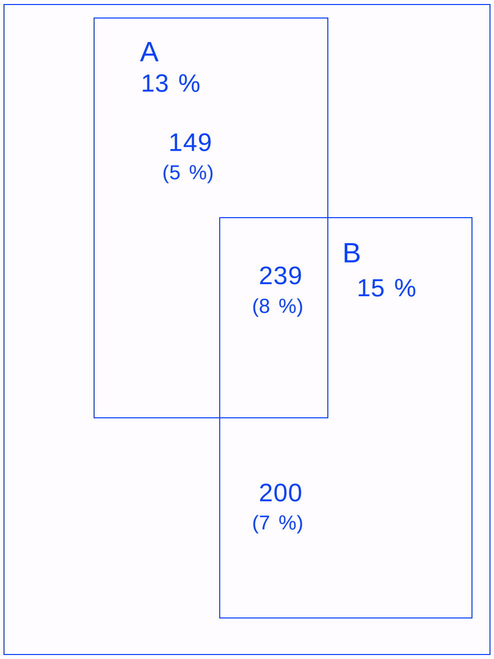
By specifying **gen(vd1)** a new variable is added to the dataset (The routine will **not** save the users file with the new variable, must be done afterwards). When creating a new variable a **note** is added to the dataset indicating the date and which variables with which values were applied. Additionally any **if/in missing** options will be added to the note. One other possibility is to enlarge the textsize by applying the general command **set textsize.** In figure 3 it was set to 120. The log file will indicate the creation of the new variable:

```
. set textsize 120
venndiag astma season in 1/3000, saving(figure3) show(xcpt)
label(tx)                            /*
>   */   r1(Allergi related symptoms) r2(A:eczema)  gen(vd1)
/*
>   */   r3(B:astma) r4(N=2948) r5(Text: set textsize 120)
r6(t1..t3 available for other texts)
__________________________________________________________________
____

Venn diagram of variables: astma season

.. output omitted .....

New variable created. Name: vd1  Label: astma season(vd)
notes added:

vd1:
  1.  generated by venndiag.ado on 18 Nov 1998 21:48 . Variables
and values were
  2.  astma:1 season:1
  3.  vd1=miss for 1000 records excluded by:in 1/3000
  4.  vd1=miss for 52 records with missing values (.)
```

figure 3 here

The label of the generated variable will indicate which variables were used for creation, such as:

```
. tab vd1

astma       |
season(vd)  |      Freq.     Percent        Cum.
------------+-----------------------------------
         -- |       2360       59.00       59.00
```



Allergi related symptoms

A:eczema

B:astma

N=2948

Text: set textsize 120

t1..t3 available for other texts

```
          A |        149        3.72       62.72
         AB |        239        5.98       68.70
          B |        200        5.00       73.70
       miss |       1052       26.30      100.00
------------+-----------------------------------
      Total |       4000      100.00
```

A is the first variable in the label, B the second etc. When tabulating the variable records omitted from the graph will be given the value ”**miss**”. Here the exclusion was caused by an **in 1/3000** (leaving 1000 records) clause and **52** records having missing values in astma and/or season variables. This information is saved with the dataset in a note as shown above and can be shown with the general command **note**

Historical note:
John Venn (1834-1923) was British. He worked at the University of Cambridge on logic and developed the “Venn Diagram” – a diagrammatic method of illustrating propositions by inclusive and exclusive circles probably during the years 1866-1881. According to “dictionary of National Biography (1922-30)”, page 869 the idea had been developed previous to the publication of Symbolic Logic in 1881, but no primary source is given. In the foreword of *Symbolic logic* John Venn states that some of the ideas had been presented earlier and in a historical chapter he writes on page 511:“*So far as I have been able to ascertain, this plan (as applied to closed figures) was first employed by Thomson in the second edition of his Laws of Thought 1849”.* The actual introduction of closed circles to represent combinations of variables therefore most likely dates back around 1850, but with John Venn deriving more stricly the relationship between logical statements and diagrammatic representations. In the original drawings Venn applied circles to represent two and three variables and elipses to represent four variables. (For reasons of programmatic simplicity squares have been chosen in the venndiag.ado routine, future enhancements could change this)

References

Venn J. Symbolic Logic. London, Macmillan & Co, 1894. (2nd revision 1894)(1st ed. 1881)

Thanks to ph.d. student MD Charlotte G.Mörtz for testing and comments.