


```

*) Letter value to trace: default median, =2 (hinges),
   3 (eights) .. 9, =0 (range)
**) By default v is broken into 3 strips
<opt> [MINFREQ=nmin] [POWER=reexp.power]
      [SORT <crit> {DES | ASCENDING}]
      [REMOVEDMEDIAN | STANDARDIZE]
<crit> [MEDIAN] | MEAN | N | MIDSPREAD | SDEVIATION
      | RANGE | VAR=critvar#
<dopt> See ?DLINE CODED for valid options

```

The Analysis of groups (I)

1 Prelude: Contrasts and systematic comparison

Understanding a data set and relationships between variables can often be gained from simple tools combined with high interactivity. This is a recurrent theme in exploration. A good example of this are coded lists.

Detailed insight can be obtained from studying contrasts between groups (classes) of observations. The selection mechanism provided by the EDA Software offers a very powerful tool for this kind of analysis. Consider the following commented command sequence as an illustration of what you might do. Please note that you can use any analysis command between the selection commands (here the LIST command hints a some meaningful collection of variables)...

```

>INCLUDE #GNPAgr>20           ! Countries with agric gt 20% of GNP
>LIST ....
>ELSE                         ! ... want to see all OTHER countries
>LIST ....
>REMOVEIF #GNPIND>35         ! remove those with GNPInd gt 35%
>LIST ....
>ELSE                         ! let's see the others (unselected)
>LIST
>ELSE                         ! come back to the previous countries
>LIST ....                   ! want to see'em again
>OR #GNPServ<20              ! add countries where GNPServ > 20%
>LIST
(....)

```

Make sure to understand that the various commands change the selection by adding or removing observations from the current selection or by switching from the currently selected cases to the unselected cases (ELSE command). Note that the LIST command here represents any number of analysis commands.

2 Introduction to groups

Groups (classes) occur in many situations:

- Natural classifications, e.g. continents
- Categorical data
- Theoretically defined groups/classes
- Classification procedures
- Breaking/slicing variables

Classification is omni-present, even in our everyday thinking; results based on the comparison of groups are often easier to use and to communicate, than results from other (statistical) tools.

Often we will want to define groups by breaking quantitative information into groups in order to clarify and simplify a relationship (categorization). The result of any group defining procedure, simple or complex, is a categorical variable showing the group to which an observation belongs. In many situations this variable is *nominal*, but sometimes *ordinal* as in the case of interval or fractional coding, where the groups reflect some order.

In the EDA software there is a special “variable” attached to a WA, called a GVAR defining group membership. By default no GVAR exists, i.e. no groups are defined. There are many different tools defining a GVAR. Whenever a GVAR is defined it is saved together with the WA (PUT) and restored when you GET that WA. The GVAR contains integer values. A membership number of zero means that the observation does not belong to any group.⁵

A GVAR may be defined by a number of commands. The CODE command may be used to build a GVAR by cutting interval variables into groups using the same rules we have used for coding. Other commands defining GVARs include HISTOGRAM, FREQUENCY, BREAK and others, and of course, as we will see later, cluster analysis (HIERARCHY and CLUSTER). It is also possible to store any suitable variable as a GVAR.

Many commands use and show the GVAR:

- ◆ Most plotting and listing facilities (case identifiers are often shown together with the group memberships).
- ◆ The case selection mechanism may be used to analyze groups separately (see the ANALYSE command).
- ◆ Many commands have a GVAR option (e.g. PLOT, HISTOGRAM) or a BYGVAR option (e.g. STEMLEAF, LIST).

A number of specific commands are designed for group analysis and definition:

GANALYSIS	Group analysis (coded displays)
GSUMMARY	Group summaries (numerical displays)
TRACES	Parallel boxplots for each group

GVAR manipulation

GVAR	Manipulate GVAR
MEMBER	Display group memberships
CODE GVAR	Define a GVAR by coding
ANALYSE	Case selection: analyze a group

Other

FREQ BYGVAR	Frequency by groups
STEMLEAF BYGVAR	Groupwise stem and leaf
DISPLAY BYGVAT	Stat. summaries for groups

2.1 GANALYSIS, GSUMMARY

The GANALYSIS and GSUMMARY commands require a GVAR⁶ The GANALYSIS command shows numerical summaries for each group, as well as for the overall variable.

```
>GSUMMARY #INCOME-#FROST
```

5. Note that in most situations “group” 0 is not shown or is shown as blank.

6. or alternatively a suitable variable in the WA.

```

50 states
Groups defined by GVAR:divison
Group summaries: median & midspread (above/below)

```

Variable	East	S.Cent.	N.Cent.	West	Overall
Income	4558.00	3848.00	4594.50	4660.00	4519.00
	622.00	824.00	250.00	616.00	832.00
Ill	1.10	1.75	0.70	0.60	0.95
	0.40	0.80	0.20	0.90	1.00
LifeExp	71.23	70.07	72.28	71.71	70.68
	1.28	1.54	1.83	1.74	1.79
Murder	3.30	10.85	3.75	6.80	6.85
	2.40	3.35	6.05	4.70	6.40
HS	54.70	41.70	53.25	62.60	53.25
	4.60	9.45	5.60	4.40	11.40
Frost	127.00	67.50	133.00	126.00	114.50
	46.00	47.50	31.50	123.00	75.00
Group N	9	16	12	13	

In this example all groups are shown. Small groups (the NMIN option lets you define the minimal number of observations) are not included, as well as observations not member of any group.⁷

Furthermore if the summary does not fit on the screen (many groups and/or big numbers), you will be asked to select the groups to show.

An alternative way of displaying the same information is the *coded form* of the same table, i.e. reference coding for each group: +/- symbols indicate the distance of the group center to the overall center of the variable.

```
>GANAL #INCOME-#FROST DIV_UNIT=4
```

```

50 states
Groups defined by GVAR:divison
Coded group differences.
Each symbol corresponds to 1/4 midspread deviation from the grand median.

```

Variable	East	S.Cent.	N.Cent.	West	Overall
Income		---			4519.00
Ill		+++	-	-	0.95
LifeExp	+	-	+++	++	70.68
Murder	--	++	-		6.85
HS		----		+++	53.25
Frost		--			114.50
Group N	9	16	12	13	

An alternative form of display is shown below.

```
>GANALYSIS #GNPGrow-#Lit LONG DIV=5
```

```

183 countries
Groups defined by GVAR:Continents
Group summaries: median & midspread(Column 1 & 2)
Each symbol corresponds to 1/5 midspread deviation from the grand median

```

Continent	Members	GNPGrow	GNPCap	GNPAgr	GNPInd	GNPServ	Lit	Symbol
Asia	(1) has 39 members	GNPGrow(1)	GNPCap(2)	GNPAgr(3)	GNPInd(4)	GNPServ(5)	Lit(6)	
		3.80	1270.00	19.00	32.00	43.00	77.00	
		6.20	5980.00	28.50	20.00	15.00	32.00	++
Africa	(2) has 53 members	GNPGrow(1)	GNPCap(2)	GNPAgr(3)	GNPInd(4)	GNPServ(5)	Lit(6)	
		2.20	393.00	30.00	25.00	43.00	50.00	
		3.30	635.00	31.00	17.00	13.00	27.00	-
								++
								--

7. Seen from EDA they are member of group 0.

```

Europe ( 3) has 30 members
GNPGrow(1)      2.00      1.40 :
GNPCap(2)      12543.50  13221.00 : $+++++++
GNPAgr(3)       5.50      10.00 : --
GNPInd(4)      36.00      7.00 : ++
GNPServ(5)     59.00     12.00 : +
Lit(6)         99.00      4.00 : ++
N&C.Am ( 4) has 31 members
GNPGrow(1)      3.00      4.05 :
GNPCap(2)      2762.00   6271.50 :
GNPAgr(3)       9.00     13.50 : -
GNPInd(4)      22.00     14.00 : -
GNPServ(5)     64.00     22.00 : ++
Lit(6)         94.00     9.50 : +
(some continents not shown)

```

In fact the example below illustrates a possible pitfall of the display. As globally and within each group you find wildly different observations the overall midspread is usually quite important, i.e. even with 1/5th of a midspread as unit, the his not very much contrast to be seen. Certainly less than one would expect from the knowledge on the data. Here the next display might be helpful.

```

>GANALYSIS #GNPGrow-#Lit LONG BOXPLOT

183 countries
Groups defined by GVAR:Continents
Group summaries: median & midspread(Column 1 & 2)
Asia ( 1) has 39 members
GNPGrow(1)      3.8      6.2 : ° _____ ≈ _____ °
GNPCap(2)      1270.0   5980.0 : ≈ _____ °
GNPAgr(3)      19.0     28.5 : _____ ≈ _____ °
GNPInd(4)      32.0     20.0 : _____ ≈ _____ °
GNPServ(5)     43.0     15.0 : °°° ° _____ ≈ _____ °°
Lit(6)         77.0     32.0 : ° _____ ≈ _____ °
Africa ( 2) has 53 members
GNPGrow(1)      2.2      3.3 : ° ° _____ ≈ _____ ° °
GNPCap(2)      393.0    635.0 : ≈ _____ ° _____ ° ° °
GNPAgr(3)      30.0     31.0 : _____ ≈ _____ °
GNPInd(4)      25.0     17.0 : _____ ≈ _____ °
GNPServ(5)     43.0     13.0 : ° _____ ≈ _____ ° °
Lit(6)         50.0     27.0 : _____ ≈ _____ °
Europe ( 3) has 30 members
GNPGrow(1)      2.0      1.4 : ° _____ ≈ _____ ° °
GNPCap(2)      12543.5  13221.0 : _____ ≈ _____ °
GNPAgr(3)       5.5     10.0 : _____ ≈ _____ °
GNPInd(4)      36.0     7.0 : ° _____ ≈ _____ ° _____ ° °
GNPServ(5)     59.0     12.0 : ° °° ° _____ ≈ _____ °
Lit(6)         99.0     4.0 : ° _____ ° °° ° _____ ≈ _____ °
N&C.Am ( 4) has 31 members
GNPGrow(1)      3.0      4.0 : ° _____ ≈ _____ ° °
GNPCap(2)      2762.0    6271.5 : ≈ _____ ° °° °
GNPAgr(3)       9.0     13.5 : _____ ≈ _____ °
GNPInd(4)      22.0     14.0 : _____ ≈ _____ ° _____ °
GNPServ(5)     64.0     22.0 : ° ° _____ ≈ _____ °
Lit(6)         94.0     9.5 : ° ° _____ ° _____ ≈ _____ °
(South America and Aust/Oceania not shown)

```

An additional tool you might use are profiles. The PROFILE command, used in its basic form to show profiles of a single observation or several observations, may be used to show profiles of the groups or profiles of a specific group and the observations in that group. The example below illustrates this last possibility.

Each number marks the position of an observation in group #1 and the group centroid (median) is shown using the '#' symbol.

```
>PROFILE #Income-#Frost GROUP=1
```

```

50 states
Group defined is from:divison
Profile for states CT : "1" ME : "2" MA : "3" NH : "4" NJ : "5" NY : "6"
PA : "7" RI : "8" VT : "9"
Group# 1 centroid marked with #
Label min max.
Income 3098.00 2 9 4 7 # 3. 6 51 6315.00
Ill 0.50 9 $ 7 # 8 6 . 2.80
LifeExp 67.96 $6 .5 # 9$ 1 73.60
Murder 1.40 821# $ 7 . 6 15.10
HS 37.80 8 7 56. # 1 $ 3 67.30
Frost 0.00 6 .3 5 7# 1 29 4 188.00

```

Command syntax

GANALYSIS command

Displays coded group summaries.

```
GANALYSIS | [MEDIAN] | [DIV_unit=val] <options>
          | MEAN |
```

```
<options> [GVAR=var#] [NMIN=min_members] [KEEP_GROUP_0]
          [LONG {DIFFERENCES} {BOXPLOT}]
```

- Requires a GVAR or GVAR=var# option.
- If more than five groups are found, you will have to enter a selection of groups (LONG format excepted).

GSUMMARY command

Display summary statistics for groups.

```
GSUMMARY <sum> <options>
```

```
<sum> [MEDIAN] | MEAN | IFENCes | OFENCes | HINGes | RANGE
```

See also: DISPLAY BYGVAR.

```
<options> [GVAR=var#] [NMIN=min_members] [KEEP_GROUP_0]
          [LONG {DIFFERENCES} {BOXPLOT}]
```

- Requires GVAR or GVAR=var# option.
- If more than five groups are found, you will have to enter a selection of groups (LONG format excepted).

Shows profiles of cases or groups of cases
(up to 10 profiles are shown per variable)

PROFILE

```
PROFILE vlist [GROUP{=grp#}{MEAN}{NMIN=min_memb}] ["alt_sym"]
             [Case=cas#]
```

- (default) queries for casids, if you specify a single casid value is coded (in,out,far-out,adjacent)
- GROUP=gpno displays members of a single group with group centroid
- GROUPS displays group centroids for up to 10 groups

2.2 Break tables

The BREAK command produces a crosstabulation by cutting each variable into a number of intervals (default three) according to some criterion (default thirds) and counts the number of cases in each group.

The BREAK command has options to identify the observations in a particular cell and the crossclassification can of course be saved as GVAR if needed.

```
>BREAK 41 48 FIVE
```

```
157 towns
Crossbreak of class12(41) with left55(48)
Table shows counts
      c1 c2 c3 c4 c5 #
r1    1  5 25 31
r2     5  5 14  7 31
r3     5  5 13  8  1 32
r4     6 13  9  3  31
r5    20  7  5  32
#     31 30 33 30 33 157
```

In addition you may request a table containing summary information on a third variable, in the following example the medians of Labour votes in 1951.

```
>BREAK 41 48 WITH=47 FIVE
```

```
Cell contents based on left51(47) % votes for Labour and other left-wing
Table shows cell medians. Overall median=49.6
      c1  c2  c3  c4  c5
r1    43.6 36.4 32.6
r2    40.8 45.8 43.8 41.5
r3    49.5 50.2 50.0 49.5 35.8
r4    56.0 55.7 54.0 49.2
r5    64.8 58.9 58.2
```

2.3 Further details on the GVAR

It is often convenient to define a GVAR for more detailed analysis. For instance:

- ◆ Commands like the TRACE, GANALYSIS or GSUMMARY command can be used without any option
- ◆ Use a simple case selection command: ANALYSE GROUP=group# to select a particular group instead of writing a more complex expression.
- ◆ The GVAR is fully integrated and is shown automatically, without the need of options, e.g. on lists or can be added with a simple GVAR option like on PLOT 1,2 GVAR or HISTOGRAM GVAR.
- ◆ Use the MEMBERS command to show group memberships any time.
- ◆ Display group names (and not only numbers)⁸

Note that only one GVAR can be defined simultaneously. GVARs can however be copied into the WA (GVAR LOAD) as normal variables and variables in the WA can be stored as GVARs (GVAR STORE).⁹ Most commands with a GVAR option let you also specify a grouping variable in the WA instead of the currently defined GVAR; instead of writing PLOT 1 2 GVAR you will write PLOT 1 2 GVAR=#gv, where gv is a variable in the WA and taken up as a grouping variable.¹⁰

The CODE command can be used to produce GVARs by breaking variables

```
>CODE 1 GVAR NGROUPS=5
```

8. Names can be defined using the GVAR NAME DEFINE command. GVAR NAMES displays the currently defined names.

9. When storing normal variables only the integer part of a variable is considered. GVAR STORE offers a number of options for manipulation, e.g. shifting the decimal point before storing.

10. Note that the variable should contain groups; otherwise results might be rather strange. Only the integer part of a variable is considered.

breaks variable 1 into 5 groups (default three), using - of course - the usual default criterion, groups of approx. equal size. Options include equal width and more. CODE has also an interactive PLAY mode used to find an appropriate coding (try it out).

Furthermore you may refer to the GVAR in expressions:

```
>LET GVAR=(#VARX*#VARY)/100
```

computes a new GVAR (multiply VARX by VARY, divided by 100).

```
>IF GVAR > 100 THEN GVAR=0
```

Sets the GVAR to zero for all values larger than 100 (remember that '0' means no group membership

Examples (1) and (2) use a GVAR reference, i.e. refer to the GVAR as a full vector.

```
>LET G[1]=2
```

refers to the first element in G[] and sets it to 2.¹¹

Of course a reference to the GVAR is useful with more complex selections, as in

```
>INCLUDE GVAR<4 & #GNPInd>20
```

11. In fact GVAR in (1) or (2) is a short form of G[], i.e. all cases in G[]

Table des Matières

1 Prelude: Contrasts and systematic comparison	4
2 Introduction to groups	4
2.1 GANALYSIS, GSUMMARY	5
Command syntax	8
2.2 Break tables	8
2.3 Further details on the GVAR	9