PN-AAA-[illegible]
ISN 44567

62

EXERCISES IN DATA MANAGEMENT AND ANALYSIS: INFORMANT SURVEY DATA

Frank W. Young

Cornell University

With the assistance of Susan Bentley and Walter West

January, 1982

SPECIAL INSTRUCTIONS WHEN USING THE DEC

1. For these exercises, you must issue the following command from Tops-20:

@ TAKE CLASS=RS213

You may do this before or after using the otto editor, but it must be done <u>before</u> you try to run or execute your SPSS program.

The command TAKE CLASS=RS213 will give you access to the file which contains the Tunisian data. This command must be issued once, <u>each time</u> you log in.

2. Disregard all the instructions on Job Control Language (JCL) which applies to the IBM computer and not to DEC.

TABLE OF CONTENTS

Introduction

List of 80 subdistricts in the pilot zone

Map of boundaries of study areas in Central Tunisia

1. Exploring data and initial decisions

2. Histograms and regional breakdowns

3. Inspecting raw data and percentaging

4. Reliability and the computation of scores

5. Indicators constructed by the Guttman technique

6. Indicators constructed by the Z-score technique

7. Index of inequality

8. Recodes and crosstabs

9. Correlations and scattergrams

10. Transformations and special indices

Appendix: Informant questionnaire and check-list used in Central Tunisia

INTRODUCTION

This set of exercises is based on the monograph A Baseline for Rural Development for Central Tunisia (1981). That report resulted from the analysis of the data from an informant survey administered in 80 subdistricts of Central Tunisia in October, 1979. Although Baseline included both measures of rural "progress" and the partial causal analysis of these social indicators, the exercises focus on problems of measurement, particularly of the dependent variables. Put another way, the exercises attempt to move the student from the raw data set to the place where explanatory analysis can begin.

In order to make the most of these exercises, some supplementary materials are required. Included in this set of exercises are maps of Tunisia as a whole and another outline map of Central Tunisia. An extra copy of the latter is needed for one of the exercises. The serial numbers for the 80 subdistricts and their names are inserted just after the outline map. At the end of the materials, as an appendix, we have included an English translation of the questionnaire, which can serve as a code for the 354 variables on file.

The labels for these variables on the SPSS file are available and will have to be consulted for some of the exercises. Similarly, we have a complete listing of the values of all the variables, as well as frequencies and descriptive statistics. Two other separate, but necessary, items are the monograph that describes this research, "A Baseline for Rural Development in Central Tunisia" and the relevant handbooks for the SPSS programs. These exercises generally refer to the Primer, but some of the exercises depend on the general handbook, and a couple refer to the Update.

As will be evident, the exercises start from the raw data set and move in a sequence that one would normally take for this type of descriptive and applied research. It is assumed, of course, that the data have been collected

according to some systematic and reliable set of procedures. These may be described briefly as follows:

(1) Review of literature. A publication entitled The Rural Development Inventory (Young, 1977) summarizes most of the articles that have led to the particular methodology and substantive emphasis used in the Central Tunisia informant survey. Additionally, we consulted informant questionnaires that have been used in Peru, Tanzania, and the Philippines. The new element in this questionnaire is the attempt to measure poverty, agricultural productivity, and farm practices on the basis of informants' estimates rather than the usual farm or household questionnaires.

(2) Exploratory research. Most questionnaires of this type are preceded by at least a month or two of exploratory fieldwork, but that was not possible in this case. As a substitute, we interviewed members of a team of planners from Wisconsin, read all the reports that were available, and made several visits to the countryside, which included conversations with a few local leaders. The first version of the questionnaire was drafted in July 1979. This version embodied a two-part structure: Estimates of means for households are farms with respect to poverty, productivity, farm practices, and equipment, and institutional patterns (agriculture, occupational structure, marketing, administration, water sources, social life, migration, urbanization). The questionnaire was designed to serve some of the planning needs as well as to have utility for monitoring and evaluation.

(3) Revision and pretesting (August-September 1979). In August, many items were added or deleted, but a major change was the elimination of all the questions that could be better answered from data available from (government) agencies Tunisian staff made further changes, and the Director approved the final version with about three deletions. The French version of the questionnaire

was then translated into Arabic and pretested in nearby places. At this point the questions were assigned numbers.

(4) Administration of the questionnaire and field editing. The questionnaire was administered by Tunisian staff stationed in the eight delegations, with supervision and training by two Tunisian planners from the central office and an American consultant. The fieldwork, which included training sessions for approximately ten interviewers, took six weeks. The local leaders (omdahs) were usually interviewed in or near the district office, and after completing the questionnaire, one of the supervisors edited it to see that all questions were answered correctly. However, this was not possible in all cases.

(5) Coding and editing. Although the questionnaire was precoded, it was still necessary to transfer the numbers corresponding to each of the answers for the 354 questions or parts of questions to a codesheet in preparation for input to the computer. This process uncovered a number of errors, which had to be checked against the original questionnaire.

List of 80 subdistricts in the pilot zone.

District of Foussana
1) Afrane
2) Boudriass
3) Hazza
4) Foussana
5) Ouled Mahfoudh
6) Ghdira
7) Brika (el)
8) M'ziraa (el)
9) Khemouda

District of Thala
10) Thala-Est
11) Thala-Ouest
12) Dachra
13) Haidra
14) Ajred (el)
15) Tbaga
16) Mkimen (el)
17) Chaffai (ech)
18) Oueljet Edhel
19) Boulahnech
20) Hmad
21) Joui
22) Oued Erracheh
23) Zalfane
24) Sidi Shil (Ain Jadida + Bourmajna)

District of Sbeitla
25) El Athar + El Kadra
26) Echaraa
27) Edoulab
28) Karaa El Hamra
29) Errakhmet
30) Gounna (el)
31) Machrek Echames
32) M'zaraa (el)
33) Oussaia (el)
34) Semama
35) Sbeitla

Delegation of Sbiba
36) Ahouez (el)
37) Ain Zaien
38) Ain Khemaissia
39) Brahim Zahar
40) Oued Lahtab
41) Sbiba
42) Thmad

Delegation of Djedelienne
43) Ain El Hamedna
44) Ain Oum Jedour
45) Bouajer (el)
46) Brik (el)
47) Hamaima (el)
48) Krine (el)
49) Djedelienne
50) Remada
51) Terbah
52) Toucha

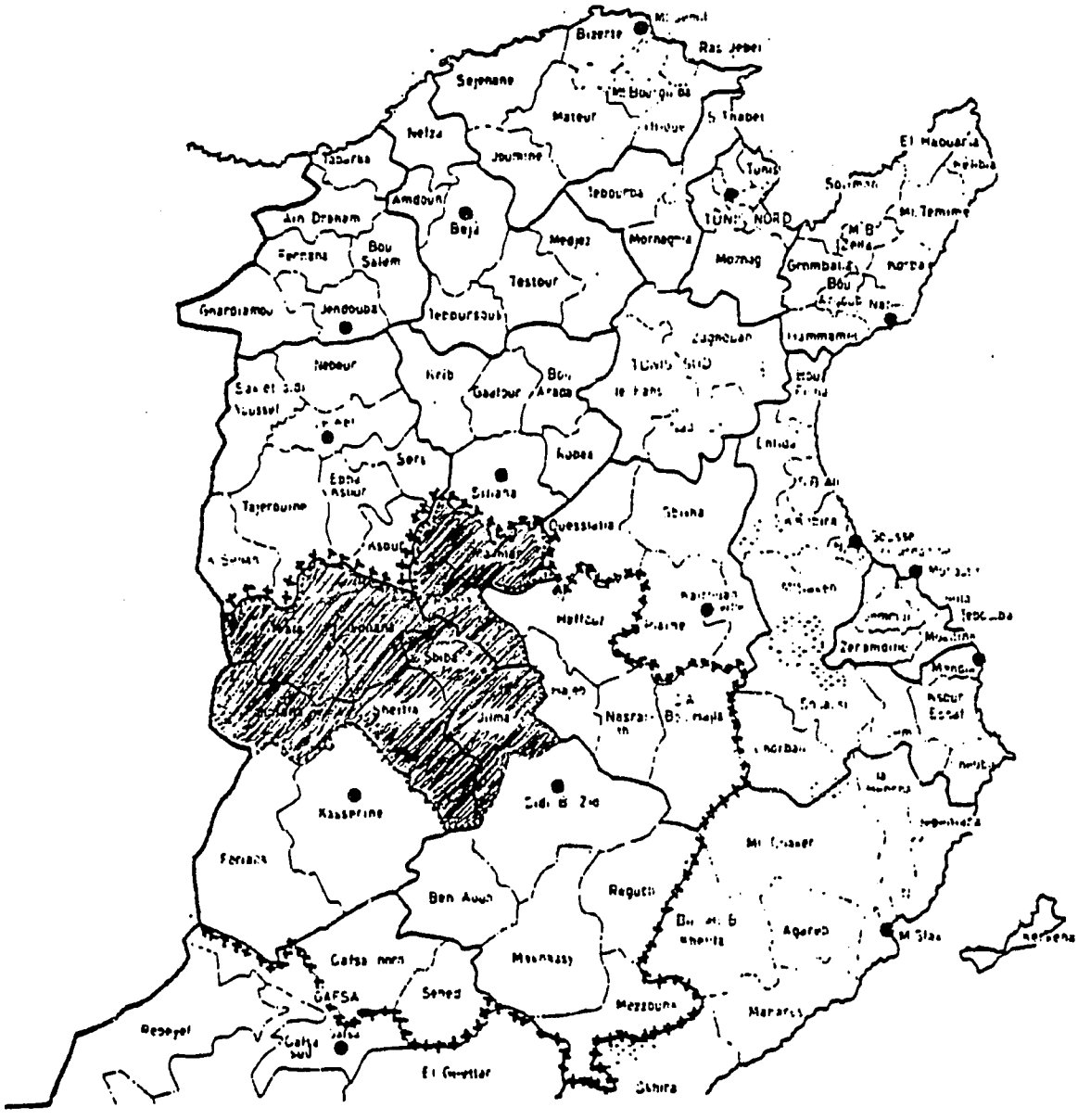
District of Rouhia
53) Haria (el)
54) Hababsa
55) Msahla (el)
56) Rouhia
57) Smirat (es)
58) Jemailet (el)

District of Makthar
59) Benihazem
60) Bez
61) Garaa (el)
62) Makthar
63) Ras El Oued
64) Sayar
65) Sadine (es)
66) Soualem (es)

District of Kesra
67) Foudhoul (el)
68) Karia (el)
69) Kesra
70) Louza (el)
71) Mansoura (el)

District of Jelma
72) Abiadh (el)
73) Amra (el)
74) Baten Laghzal
75) M'Ghila
76) Ghdir Zitouna
77) Jelma
78) Selta
79) Sabala (es)
80) Sid (es)

Map Boundaries of study areas in Central Tunisia.



TUNISIA

Study Area

key:

governorate

delegation district

governorate seat

pilot zone

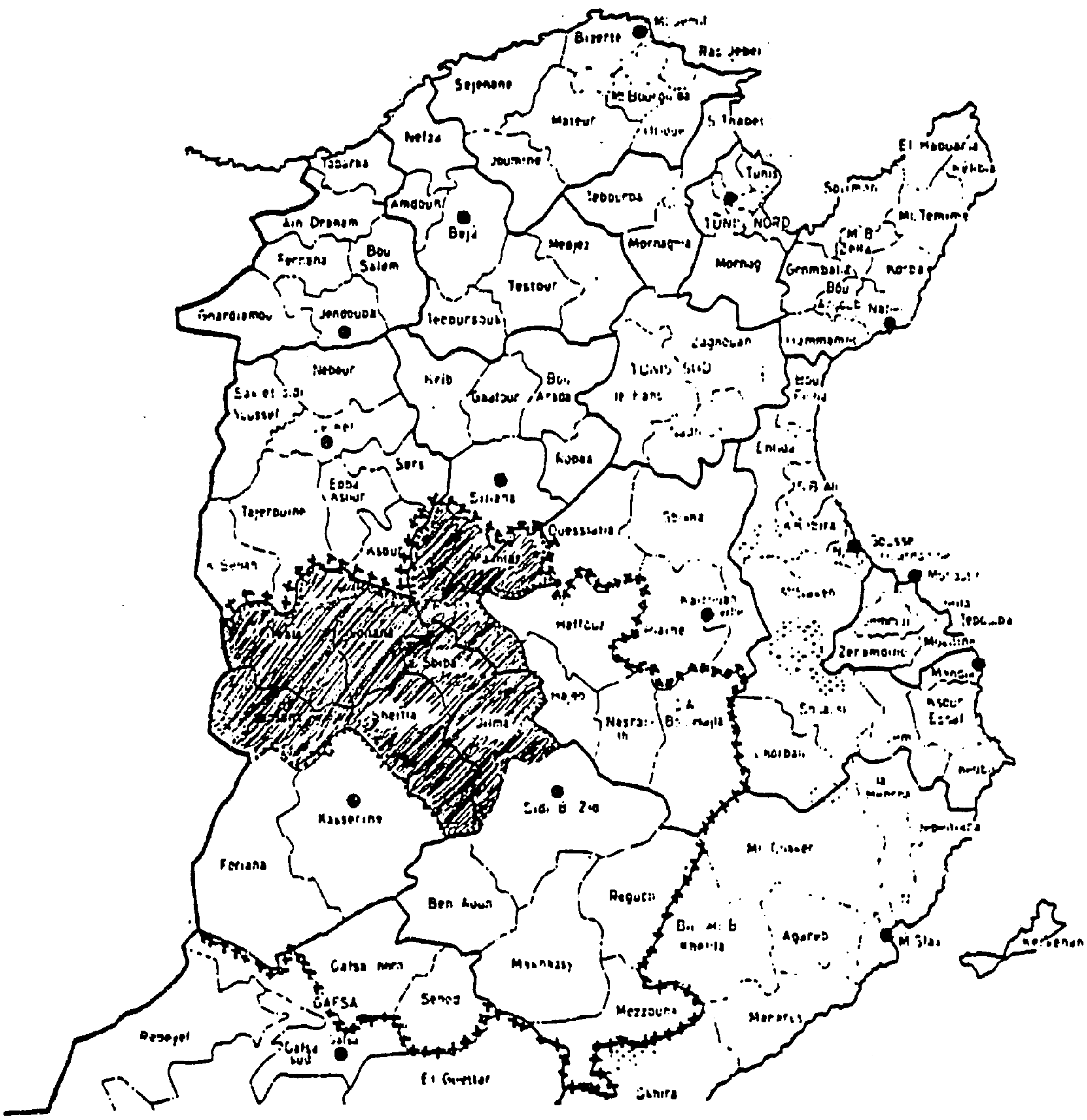
operational region

50 km

Delegations in Target Area: Makthar, Kesra, Rohia, Thala, Djilma, Djedlianne, Sbiba, Foussana, Sbeitla

Additional Delegations in ODTC Area: Maknassey, Sidi Bou Zid, Ben Aoun, Mezzouna, Ouled, Haffouz, Regueb, Feriana, Kasserine, Nasrallah, Bouhjla, El Alam, Hadjeb El Aoun, Sned, Gafsa Nord

Map Boundaries of study areas in Central Tunisia.



TUNISIA

Study Area

key:
- governorate
- delegation district
- governorate seat
- pilot zone
- operational region

50 km

Delegations in Target Area: Makthar, Kesra, Rohia, Thala, Djilma, Djedlianne, Sbiba, Foussana, Sbeitla

Additional Delegations in ODTC Area: Maknassey, Sidi Bou Zid, Ben Aoun, Mezzouna, Ouled, Haffouz, Regueb, Feriana, Kasserine, Nasrallah, Bouhjla, El Alam, Hadjeb El Aoun, Sned, Gafsa Nord

EXERCISE 1: DATA EXPLORATION AND DESCRIPTION

Problem: How does one start a data analysis, and what initial decisions must be made?

Short answer: List the cases for a familiar cluster of variables, look for extreme values or other problematic features, and then run CONDESCRIPTIVE for total sample, total sample minus the special cases, and the refined sample within subregions. The components of this general statement will be clarified below.

Reading: SSPS Manual, 137-139.

Output: Attached, pp. 1-19, plus page A. The four variables are:

V449, Total Population

V452, Population Density

V454, Number of Agglomerations

V462, Resident Population of the Communes.

Commentary: These four variables are taken from the official 1975 census and are presumed to be correct. (Ways of dealing with questionable or missing data will be taken up in another exercise.) The total resident population of the subdistrict includes all those who claim the subdistrict as their home, even if they are away working. The second variable, Population Density, was already computed and listed in the census. Density is the number of inhabitants divided by the number of square kilometers. The Number of Agglomerations refers to the subcommunities or hamlets that dot the countryside. "Agglomeration" is a French word which is still retained for these particular variables in the code, although all the others have been translated. These subcommunities represent the beginnings of "nucleated settlement," which, however, can only occur when there is enough water to support dense populations. A normal pattern in this area is for single or occasionally joint families to live scattered out across the countryside wherever they can find a little bit of water during the rainy season or perhaps a patch of arable land. (Variable 462 refers to the population of the "communes," that is, the official urban centers. This variable is important as an indicator of a special kind of population and will be considered

in more detail below.)

Steps in comprehending output:

1. Study the next page in this handout that explains the control cards, which are listed on page A, and pages 1, 5, 11, and 13. These are explained in the SPSS manual, but the TA will discuss them in detail.
2. Scan the listing of values for each of the subdistricts (cases) for the four variables, starting on page 2. You are looking for numbers that seem "odd." For example, a population of 126 for subdistrict 11 is odd.

   Looking on down one sees that subdistrict 35, Sbeitla, has a large population, high density, and all the subdistrict population is classified as communal. In contrast, subdistrict 41 has only about half of its population classified as communal. Subdistrict 62, which is the old Roman town of Maktar, has a high density, and all of its population is classified as urban.
3. Identification of Special Cases

   These urban centers, particularly the completely urban units, pose a problem for all subsequent analysis. After all, there are not going to be any farms in these centers; yet that is what the questionnaire deals with. What should one do? This issue came up in collecting the data, where there was a question about whether we should even interview the leaders in the towns. The field decision favored interviewing them, but as the work began, it was clear that they should be separated for most of the analysis. In addition to having few or no farmers, the three largest places were old Roman towns that are still important in regional economy and are probably responding to different

forces than those that impinged on the more rural subdistricts. Had there been 10 or 20 of them, they could have been separately analyzed as a subsample, but with only 3 (actually 4 subdistricts, counting the 2 in Thala) the decision was to delete them. These and the other towns are discussed in Baseline starting page 17. The map on page 15 shows the general locations of these places, and the table on page 21 details all of their characteristics. At this point you might wonder how this subtraction of 4 cases is to be accomplished. One way to do it would be to extract the 4 punched cards. However, this data came on a tape, so that was not possible. Also, we wanted to include these places for some of the preliminary descriptive analysis. So we chose the second method which is to use the *SELECT IF command in SPSS as shown on page 11 of the output.

4. Mapping: The maps for these three variables appear on pages 47-49 of Baseline. Such maps are easily made; they involve some decisions as to cutting points and choice of cross-hatching (which can be purchased commercially) and in the case of Map 2.8, it was necessary to label the geographic location of subcommunities as "approximate," because they were not precisely identified in space. Note, by the way, how the three towns, soon to be deleted, show up on Map 2.6.

5. Descriptive Statistics for the Total Sample

a. The descriptive statistics for the total sample are shown on page 4.

b. The CONDESCRIPTIVE program routinely calculates the ten statistics that are listed for each variable, but only several of these are relevant at this point. The first of these is the mean of the total population, which is 2904.937. In practice, you would round that off to 2905. You can verify this figure by dividing the SUM of 232395. by 80. More on SUM later.

Most means are fairly straightforward, but the mean for V452 Population Density, requires some explanation. That mean was calculated in this case by adding all the density figures and dividing by 80. However, one could have done it differently. A weighted mean" (by population) is caculated by summing the populations (as in SUM), and dividing by the number of square kilometers in the 80 subdistricts. One is often presented with a choice of this type, and the rule is to use the calculation that is appropriate for the unit of analysis. In this case we are interested in the subdistricts, not the individual people. So the unweighted mean of 37.925 is appropriate. Note that this problem arises only when you are dealing with ratios. It is not a problem for V449 or V454.

c. The next number to look at is the Range, which is the difference between the Minimum and the Maximum. You already know that the minimum population will be 126, because that number turned up in the initial scan. Similarly, the minimum for population density is 13, with the maximum of 100, which occurred twice in the two large towns. The range for number of subcommunities goes from 0 to 9, and in this case it is important to know how many "0" subdistricts there are. The only way (at this point) to find that out is to go back to the listing of cases. In social research there is something about "0" that sets it off from all other values, so "0" cases should be investigated in detail.

d. The Sum is just what it says, the sum of all the values. It is included in CONDESCRIPTIVE as a way of checking coding and punching. when it is done from, say, census sheets. The census shows the total population for larger units, and it should be identical with this figure.

e. The last number that you need to consider is the standard deviation, which is a measure of dispersion. The formula is given in the SPSS manual, but for our purposes you need to know that 68.26 percent of the population means fall between 2,904 and $\pm$1,414.610, that is, between 1,490 and 4,319. With a little practice, this measure can prove its usefulness.

6. Subregions as Basis for Description

For many purposes, describing the variables for the whole sample is sufficient; but for rural development work, most description should be refined by computing the same variables within subregions. The basis for delineating subregions is explained starting page 28 in Baseline, and the general goal is illustrated by Table 2.10 on page 46. Briefly, you

should use the official administrative units, in this case the nine districts, as your building blocks; require your subregions to consist of contiguous districts; try to take ecological homogeneity and historical associations into account, but look at a range of variables, including cultural patterns; and finally, if you have reason to believe that districts within a given subregion will show contrast, then organize the subfile so that the figures for the separate districts can be shown. Such region/contrasts within are illustrated in Table 2.5 on page 31 and in most of the tables following it. However, for most purposes, it appeared that only the Southeast (composed of Sbeitla and Jelma) typically showed internal contrast, so only that subregion was divided up in defining the "subfiles" that are listed on page 1 of the output and actually implemented on page 5. Page 6 shows the result for the Western subregion; N=80 for the whole file, and there are 24 cases in the Western subfile. Pages 7-10 show the descriptive statistics for the other subregions and the districts comprising the southeastern subregion.

Effect of Selecting Out Special Cases

Earlier on, we noted that the three towns (actually four subdistricts) should be eliminated from this sample because they were different in kind from the rest of the subdistricts. Such deletion is easily accomplished by the *SELECT IF command as shown on page 11. Page 12 recalculates the descriptive statistics for the four variables under study with the new, reduced sample of 76. This output should be compared to its counterpart on page 4. Page 13 then calculates the descriptive statistics for each subfile, deleting the four cases. For example, page 14 shows what happens to the western subregion when the deletions are made. The two subdistricts that comprise the town of Thala are dropped out and the number of subdistricts is reduced to 22. Compare this output with its counterpart on page 6. And in general, study all the rest of the outputs and understand where they fit. Notice also that this whole operation costs $2.36.

8. To sum up, one goal of the initial exploration is to make basic decisions on how the rest of the analysis will be organized. Two such decisions pertain to the elimination of special cases, (in this instance, the three towns) and the decision on subregions. Both of these require a lot of background information, many preliminary comparisons of pertinent variables, and some general conception of where the analysis is going. You should take your time and discuss it on all sides, because a mistake at this juncture will be relatively costly to correct later on.

Once you have made these decisions you are ready to list out all the variables. If all is well with these (and you have handled the missing values correctly), you can run CONDESCRIPTIVE and FREQUENCIES, as appropriate. However, these have already been done for this data set, and you can find them in the "Data Book" available in Warren 34.

CONTROL CARDS:

The following represent the cards you will need to obtain a basic description of the on-line data set and Exercise One. The numbers below correspond to those on the output sheets.

(1) Job card
This card identifies you and your account number to the computer.
You will be using a HASP job card for these programs. Columns 16-22 are for your access code, columns 25-44 are for your name, column 52 is for the job priority code, and columns 78-80 are for the destination of the printout.

(2) /*LIM REGION=19ØK (Note; spaces must appear as shown.)
This card tells the computer to allocate a specific amount of memory for your job. Start the card in Column 1.

(3) // EXEC SPSS
This card tells the computer to go to the disk library and find the catalogued set of procedures you will run (i.e., SPSS).

(4) //FTØ3FØØ1 DD UNIT=SYSDA,DSN=Y8H.RSOC,DISP=SHR (Note: "Ø" is a zero.)
This card tells the system that the data set you are working with resides on a disk. The data set is catalogued under the name Y8H.RSOC and must be specified in exactly that manner or the job won't run. Make sure DISP=SHR. It will allow others to access the data at the same time - thus speeding turn-around time for everyone.

The cards that follow these four are the actual SPSS program which the computer already has. You may ignore these for the time being; however, error messages are contained here, often in the beginning as ABEND=... or in a statement at the end of the computer jargon.

```
1               16
(5) GET FILE        BIGTUN.FIL
```

This card must be prepared in just this manner; without it your job won't run. This card simply tells the system to locate the SPSS file "RSOC" so that you can work on it. It generates the information in the next printed lines. Note that the format for all cards is keyword in columns 1-15 and text in columns 16-80. Also, the "card" number refers to the number of the command, not the actual number of cards, which corresponds to the separate lines.

(6) LIST CASES CASES=80/VARIABLES=V449,V452,V454,V462
This card tells the computer to list the four variables by case for all 80 cases.

(7) TASK NAME CONDESCRIPTIVE OF POP VARIABLES FOR ALL CASES
This names the following task for easier identification. This card is optional.

(8) CONDESCRIPTIVE V449,V452,V454,V462
(9) STATISTICS ALL
These two cards name the procedure to be run and the statistics to be computed for the list of variables, in this case V449,V452,V454, and V462.

(10) TASK NAME CONDESCRIPTIVE BY SUBFILE

(11) RUN SUBFILES EACH
This card tells the computer to carry out the next procedure separately for each subfile.

(12) CONDESCRIPTIVE V449,V452,V454,V462

(13) STATISTICS ALL

(14) TASK NAME CONDESCRIPTIVE WITHOUT FOUR CITY CASES

(15) *SELECT IF (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
This card tells the computer to select those cases where ID does not equal 102002 or 102002 or 103011 or 207004. The starred SELECT IF indicates that this condition is valid for the next procedure only.

(16) RUN SUBFILES ALL
This card tells the computer that all the subfiles should be grouped together for the following procedures. This command invalidates command number 11 unless it is given again later in the program.

(17) CONDESCRIPTIVE V449,V452,V454,V462

(18) STATISTICS ALL

(19) TASK NAME CONDESCRIPTIVE WITHOUT FOUR CITY CASES BY SUBFILE

(20) *SELECT IF (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))

(21) RUN SUBFILES EACH

(22) CONDESCRIPTIVE V449,V452,V454,V462

(23) STATISTICS ALL

(24) //
This card tells the computer that this is the end of the program.

```
*****************SYSMSG   ***********RM3.RD1 ***FMT <00>* *****190 LINES *** 11:09.42PM  16 NOV 81  WARREN          SUSAN BENTLEY
B | STEP NAME                    SPSS    |
B | CPU SECONDS                     2.56 |
B | CPU AU'S (CPU*5.00)            12.80 |
B | TOTAL I/O TIME, SECS           32.44 |
B | REGION USED                     152K |
B ***************************************
B * BEST CLASS FOR JOB- A *|
B *                        *|
B * AMOUNT SAVED BY        *|
B * RUNNING IN THIS CLASS  *|
B * WOULD BE:        $.01  *|
B ***************************
```

PG. A

	USED	LIMIT		MIN:SEC USED	MIN:SEC LIMIT	$CHARGED	REGION		TOTAL EXECUTION CHARGE
LINES	675	3000					MAX USED	152K	
CARDS	0	500	CPU	:03	2:00	$.25	LIMIT	160K	
FILING	0	500	I/O	:32	5:00	$.63	CHARGE	$.90	$1.78

```
B JOB 0570-EM119     FROM RM3.RD1    12:59.50 MONDAY NOVEMBER 16, 1981
B                    //EM119    JOB Y8H$$$$,'SUSAN BENTLEY        ',PRTY=4                    DLVR=WRN
C                    /*LIMITS REGION=190K
C                    // EXEC SPSS
C                    //FT03F001  DD  UNIT=SYSDA,DISP=SHR,DSN=Y8H.RSOC
C                      INPUT STREAM DATASET OF 19 CARDS
C                    //

B *-----------------------------------ACCOUNTING CONTROL RECORD (ACR) DISPLAY-----------------------------------*
B | Y8H   R S 213 NW                ACCOUNT NO.:  U964F7AA99    STATUS:   ACTIVE      ACCT BALANCE:     $179.58   |
C | DFLT RETPD: SYS        DSACCESS: SYS  DASD+CF PER DAY CHG:   $.13  DASD+CF TOT CHG:      $17.46            |
C | HASP TOTAL CHG:      $98.02                HASP CHG LIMIT:     $300.00                HASP BALANCE:     $201.98   |
C *-------------------------------------------------------------------------------------------------------------*
C //EM119    JOB Y8H$$$$,'SUSAN BENTLEY        ',PRTY=4                                    DLVR=WRN
C // EXEC SPSS
C XXSPSS PROC SPACE=20K,TYPE1=000,TYPE2=2012,TYPE3=TRK,                                    00000010
C XX      MAIN1=250,MAIN2=750,MAIN3=80,MAIN4=2250,                                         00000020
C XX      SEC1=50,SEC2=190,LIB='SYS2.SPSS.LOAD',VERSION=M,                                 00000030
C XX      ARCHIVE='DUMMY,'                                                                 00000040
C ***                                                                                      00000040
C *** CURRENTLY ACCESSING THE INSTALLED VERSION OF SPSS 8.1 (05/20/80)*** 00000050
C *** PROBLEMS SHOULD BE REPORTED TO USER SERVICES AT 607-256-4981.   *** 00000060
C ***                                                                 *** 00000070
C *** PROCEDURE   DEFAULTS  ALLOW   SUFFICIENT   SCRATCH  SPACE   FOR *** 00000080
C *** PROCESSING UP TO  1500 CASES WITH 250 VARIABLES,  OR MORE CASES *** 00000090
C *** WITH  FEWER VARIABLES.   IF  MORE SPACE  IS  NEEDED, USERS  MAY *** 00000100
C *** INCREASE IT ON THE FT01, FT02, &  SORTWK JCL CARDS BY USING THE *** 00000110
C *** MAIN1, MAIN2, MAIN3, SEC1, SEC2, & TYPE3 SUBSTITUTION PARAMETERS*** 00000120
C ***                                                                 *** 00000130
C *** IF 'GET ARCHIVE', 'SAVE ARCHIVE', OR 'MERGE FILES' IS TO BE USED*** 00000140
C *** THEN CODE ARCHIVE=  ON THE EXEC CARD IN ORDER TO ACTIVATE THE   *** 00000150
C *** NECESSARY SCRATCH SPACE.                                        *** 00000160
C ***                                                                 *** 00000170
C *** FOR THE MAXI VERSION OF SPSS CODE THE FOLLOWING ON THE EXEC CARD*** 00000180
C ***                    // EXEC SPSS,VERSION=M                      *** 00000190
C XXSPSS     EXEC PGM=SPSS&VERSION,PARM=&SPACE                        *** 00000200
C IEF653I SUBSTITUTION JCL - PGM=SPSSM,PARM=20K                                            00000210
C XXSTEPLIB  DD DSN=&LIB,DISP=SHR                                                          00000220
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSS.LOAD,DISP=SHR
C XXFT01F001 DD UNIT=SYSSC,SPACE=(&TYPE1,(&MAIN1,&SEC1)),DSN=&&LABELS,                     00000230
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(000,(250,50)),DSN=&LABELS,
C XX            DISP=(,DELETE)                                                             00000240
C XXFT02F001 DD UNIT=SYSSC,SPACE=(&TYPE2,(&MAIN2,&SEC2)),DSN=&&DATASPSS,                   00000250
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(2012,(750,190)),DSN=&DATASPSS,
C XX            DISP=(,DELETE)                                                             00000260
C XXFT05F001 DD DDNAME=SYSIN                                                               00000270
C XXFT06F001 DD SYSOUT=A                                                                   00000280
C XXFT10F001 DD DSN=SYS2.SPSS&VERSION..STAT,DISP=SHR                                       00000290
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSSM.STAT,DISP=SHR
```

```
RANSPACE REQUIRED..      100 BYTES
     1 TRANSFORMATIONS
     0 RECODE VALUES + LAG VARIABLES
    16 IF/COMPUTE OPERATIONS

PU TIME REQUIRED..    0.32 SECONDS


          4 FINISH        (SPSS GENERATED)

    NORMAL END OF JOB.
        4 CONTROL CARDS WERE PROCESSED.
        0 ERRORS WERE DETECTED.
```

SPSS FOR OS/360, VERSION II, RELEASE 8.1, MAY 20, 1980

CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM

ORDER FROM MCGRAW-HILL: SPSS, 2ND ED. (PRINCIPAL TEXT)
SPSS PRIMER (BRIEF INTRO TO SPSS)
SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)

ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS
SPSS POCKET GUIDE, RELEASE 8
KEYWORDS: THE SPSS INC. NEWSLETTER

```
DEFAULT SPACE ALLOCATION..      ALLOWS FOR..     25 TRANSFORMATIONS
WORKSPACE      17920 BYTES                      102 RECODE VALUES + LAG VARIABLES
TRANSPACE       2560 BYTES                      412 IF/COMPUTE OPERATIONS

        1 GET FILE      RSUC


                        FILE RSUC     HAS 409 VARIABLES

                        THE SUBFILES ARE..

                                   NO OF
                        NAME       CASES

                        WEST         24
                        SBCI         11
                        CNTR         17
                        NRTH         19
                        JELP          9


CPU TIME REQUIRED..     0.57 SECONDS


        2 LIST CASES      CASES=80/VARIABLES=V449,V452,V454,V462
        3 TASK NAME       CONDESCRIPTIVE OF POP VARIABLES FOR ALL CASES
        4 CONDESCRIPTIVE  V449,V452,V454,V462

***** GIVEN WORKSPACE ALLOWS FOR  358 VARIABLES FOR CONDESCRIPTIVE PROBLEM *****

        5 STATISTICS      ALL
```

CASE-N	V449	V452	V454	V462
1	2146.00	25.	2.	0.
2	2330.00	13.	1.	0.
3	2185.00	24.	0.	0.
4	3912.00	52.	3.	0.
5	3204.00	38.	3.	0.
6	2082.00	32.	1.	0.
7	2407.00	24.	0.	0.
8	2535.00	34.	1.	0.
9	4292.00	50.	0.	0.
10	8777.00	52.	0.	8700.
11	126.00	42.	0.	0.
12	2797.00	33.	1.	0.
13	4172.00	19.	1.	0.
14	1226.00	13.	1.	0.
15	1397.00	23.	0.	0.
16	1552.00	21.	0.	0.
17	1510.00	18.	0.	0.
18	3146.00	52.	1.	0.
19	3485.00	30.	0.	0.
20	1926.00	13.	0.	0.
21	1827.00	19.	0.	0.
22	3101.00	42.	0.	0.
23	2293.00	45.	0.	0.
24	4516.00	58.	0.	0.
25	4298.00	35.	0.	0.
26	2408.00	27.	1.	0.
27	2201.00	19.	0.	0.
28	3890.00	56.	0.	0.
29	1456.00	32.	0.	0.
30	2388.00	27.	0.	0.
31	3178.00	26.	0.	0.
32	2403.00	24.	0.	0.
33	2883.00	20.	0.	0.
34	2062.00	18.	0.	0.
35	8039.00	100.	0.	8039.
36	3862.00	47.	3.	0.
37	1673.00	28.	0.	0.
38	3704.00	57.	1.	0.
39	3462.00	57.	0.	0.
40	2407.00	50.	0.	0.
41	4891.00	59.	0.	1944.
42	2005.00	25.	0.	0.
43	2813.00	35.	0.	0.
44	2553.00	20.	0.	0.
45	1667.00	26.	0.	0.

21

CASE-N	V449	V452	V454	V462
46	1489.00	32.	0.	0.
47	1633.00	41.	1.	0.
48	2275.00	39.	0.	0.
49	3441.00	76.	1.	0.
50	1753.00	28.	0.	0.
51	1948.00	28.	0.	0.
52	2309.00	18.	0.	0.
53	2398.00	37.	7.	0.
54	4965.00	29.	7.	0.
55	1035.00	16.	1.	0.
56	4592.00	38.	3.	1161.
57	2444.00	26.	2.	0.
58	1608.00	13.	1.	0.
59	2050.00	56.	4.	0.
60	4725.00	54.	3.	0.
61	2510.00	33.	5.	0.
62	6068.00	100.	0.	6060.
63	1905.00	72.	7.	0.
64	3691.00	74.	7.	0.
65	2286.00	53.	4.	0.
66	3132.00	39.	6.	0.
67	1620.00	33.	6.	0.
68	2243.00	40.	5.	0.
69	3895.00	43.	4.	0.
70	2231.00	32.	2.	0.
71	4131.00	43.	9.	0.
72	3306.00	39.	0.	0.
73	2307.00	37.	0.	0.
74	2834.00	25.	0.	0.
75	4948.00	19.	0.	0.
76	2406.00	57.	3.	0.
77	4449.00	75.	3.	0.
78	4949.00	25.	0.	0.
79	1872.00	59.	2.	0.
80	2674.00	4[illegible].	2.	0.

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	2904.937	STD ERROR	158.158	STD DEV	1414.610
VARIANCE	**********	KURTOSIS	4.423	SKEWNESS	1.643
RANGE	8651.000	MINIMUM	126.000	MAXIMUM	8777.000
SUM	232395.000				

VALID OBSERVATIONS - 80 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	37.925	STD ERROR	2.058	STD DEV	18.405
VARIANCE	338.754	KURTOSIS	1.812	SKEWNESS	1.221
RANGE	87.000	MINIMUM	13.000	MAXIMUM	100.000
SUM	3034.000				

VALID OBSERVATIONS - 80 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 NAGGLOMERATIONS

MEAN	1.437	STD ERROR	0.244	STD DEV	2.187
VARIANCE	4.781	KURTOSIS	2.048	SKEWNESS	1.677
RANGE	9.000	MINIMUM	0.0	MAXIMUM	9.000
SUM	115.000				

VALID OBSERVATIONS - 80 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	324.150	STD ERROR	165.881	STD DEV	1483.684
VARIANCE	**********	KURTOSIS	23.829	SKEWNESS	4.924
RANGE	8700.000	MINIMUM	0.0	MAXIMUM	8700.000
SUM	25932.000				

VALID OBSERVATIONS - 80 MISSING OBSERVATIONS - 0

CPU TIME REQUIRED.. 0.52 SECONDS

6 TASK NAME CONDESCRIPTIVE BY SUBFILE
7 RUN SUBFILES EACH
8 CONDESCRIPTIVE V449,V452,V454,V462

***** GIVEN WORKSPACE ALLOWS FOR 356 VARIABLES FOR CONDESCRIPTIVE PROBLEM *****

9 STATISTICS ALL

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	2789.583	STD ERROR	338.906	STD DEV	1660.293
VARIANCE	**********	KURTOSIS	6.759	SKEWNESS	2.028
RANGE	8651.000	MINIMUM	126.000	MAXIMUM	8777.000
SUM	66950.000				

VALID OBSERVATIONS - 24 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	32.375	STD ERROR	2.837	STD DEV	13.897
VARIANCE	193.114	KURTOSIS	-1.187	SKEWNESS	0.350
RANGE	45.000	MINIMUM	13.000	MAXIMUM	58.000
SUM	777.000				

VALID OBSERVATIONS - 24 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 #AGGLOMERATIONS

MEAN	0.625	STD ERROR	0.189	STD DEV	0.924
VARIANCE	0.853	KURTOSIS	1.954	SKEWNESS	1.585
RANGE	3.000	MINIMUM	0.0	MAXIMUM	3.000
SUM	15.000				

VALID OBSERVATIONS - 24 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	362.500	STD ERROR	362.500	STD DEV	1775.880
VARIANCE	**********	KURTOSIS	24.000	SKEWNESS	4.899
RANGE	8700.000	MINIMUM	0.0	MAXIMUM	8700.000
SUM	8700.000				

VALID OBSERVATIONS - 24 MISSING OBSERVATIONS - 0

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	3201.273	STD ERROR	543.274	STD DEV	1801.835
VARIANCE	**********	KURTOSIS	5.619	SKEWNESS	2.204
RANGE	6583.000	MINIMUM	1456.000	MAXIMUM	6039.000
SUM	35214.000				

VALID OBSERVATIONS - 11 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	34.909	STD ERROR	7.246	STD DEV	24.031
VARIANCE	577.491	KURTOSIS	5.933	SKEWNESS	2.378
RANGE	82.000	MINIMUM	18.000	MAXIMUM	100.000
SUM	384.000				

VALID OBSERVATIONS - 11 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 MAGGLOMERATIONS

MEAN	0.091	STD ERROR	0.091	STD DEV	0.302
VARIANCE	0.091	KURTOSIS	11.000	SKEWNESS	3.317
RANGE	1.000	MINIMUM	0.0	MAXIMUM	1.000
SUM	1.000				

VALID OBSERVATIONS - 11 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	730.818	STD ERROR	730.818	STD DEV	2423.850
VARIANCE	**********	KURTOSIS	11.000	SKEWNESS	3.317
RANGE	8039.000	MINIMUM	0.0	MAXIMUM	8039.000
SUM	8039.000				

VALID OBSERVATIONS - 11 MISSING OBSERVATIONS - 0

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	2584.176	STD ERROR	236.226	STD DEV	973.986
VARIANCE	948648.779	KURTOSIS	0.196	SKEWNESS	0.944
RANGE	3402.000	MINIMUM	1489.000	MAXIMUM	4891.000
SUM	43931.000				

VALID OBSERVATIONS - 17 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	39.118	STD ERROR	3.930	STD DEV	16.205
VARIANCE	262.610	KURTOSIS	-0.119	SKEWNESS	0.753
RANGE	58.000	MINIMUM	18.000	MAXIMUM	76.000
SUM	665.000				

VALID OBSERVATIONS - 17 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 NAGGLOMERATIONS

MEAN	0.353	STD ERROR	0.191	STD DEV	0.786
VARIANCE	0.618	KURTOSIS	8.237	SKEWNESS	2.742
RANGE	3.000	MINIMUM	0.0	MAXIMUM	3.000
SUM	6.000				

VALID OBSERVATIONS - 17 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	114.353	STD ERROR	114.353	STD DEV	471.489
VARIANCE	222302.118	KURTOSIS	17.000	SKEWNESS	4.123
RANGE	1944.000	MINIMUM	0.0	MAXIMUM	1944.000
SUM	1944.000				

VALID OBSERVATIONS - 17 MISSING OBSERVATIONS - 0

27

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	3022.368	STD ERROR	315.163	STD DEV	1373.765
VARIANCE	**********	KURTOSIS	-0.413	SKEWNESS	0.684
RANGE	5035.000	MINIMUM	1033.000	MAXIMUM	6068.000
SUM	57425.000				

VALID OBSERVATIONS - 19 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	43.737	STD ERROR	4.822	STD DEV	21.018
VARIANCE	441.760	KURTOSIS	1.684	SKEWNESS	1.125
RANGE	87.000	MINIMUM	13.000	MAXIMUM	100.000
SUM	831.000				

VALID OBSERVATIONS - 19 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 #AGGLOMERATIONS

MEAN	4.368	STD ERROR	0.573	STD DEV	2.499
VARIANCE	6.246	KURTOSIS	-0.871	SKEWNESS	-0.041
RANGE	9.000	MINIMUM	0.0	MAXIMUM	9.000
SUM	83.000				

VALID OBSERVATIONS - 19 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	381.526	STD ERROR	321.954	STD DEV	1403.363
VARIANCE	**********	KURTOSIS	17.365	SKEWNESS	4.121
RANGE	6068.000	MINIMUM	0.0	MAXIMUM	6068.000
SUM	7249.000				

VALID OBSERVATIONS - 19 MISSING OBSERVATIONS - 0

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	3208.333	STD ERROR	352.527	STD DEV	1057.582
VARIANCE	**********	KURTOSIS	-0.997	SKEWNESS	0.520
RANGE	3126.000	MINIMUM	1822.000	MAXIMUM	4948.000
SUM	28875.000				

VALID OBSERVATIONS - 9 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	41.889	STD ERROR	6.174	STD DEV	18.523
VARIANCE	343.111	KURTOSIS	-0.527	SKEWNESS	0.581
RANGE	56.000	MINIMUM	19.000	MAXIMUM	75.000
SUM	377.000				

VALID OBSERVATIONS - 9 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 #AGGLOMERATIONS

MEAN	1.111	STD ERROR	0.455	STD DEV	1.364
VARIANCE	1.861	KURTOSIS	-1.917	SKEWNESS	0.508
RANGE	3.000	MINIMUM	0.0	MAXIMUM	3.000
SUM	10.000				

VALID OBSERVATIONS - 9 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	0.0	STD ERROR	0.0	STD DEV	0.0
VARIANCE	0.0	KURTOSIS	0.0	SKEWNESS	0.0
RANGE	0.0	MINIMUM	0.0	MAXIMUM	0.0
SUM	0.0				

VALID OBSERVATIONS - 9 MISSING OBSERVATIONS - 0

CPU TIME REQUIRED.. 0.42 SECONDS

```
10 TASK NAME        CONDESCRIPTIVE WITHOUT FOUR CITY CASES
11 *SELECT IF       (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
12 RUN SUBFILES     ALL
13 CONDESCRIPTIVE V449,V452,V454,V462
```

***** GIVEN WORKSPACE ALLOWS FOR 358 VARIABLES FOR CONDESCRIPTIVE PROBLEM *****

```
14 STATISTICS       ALL
```

SUBFILE WEST SBEI CNTR MPTH JCAD

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	2755.066	STD ERROR	116.840	STD DEV	1018.589
VARIANCE	**********	KURTOSIS	-0.651	SKEWNESS	0.590
RANGE	3932.000	MINIMUM	1033.000	MAXIMUM	4965.000
SUM	209305.000				

VALID OBSERVATIONS - 76 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	36.053	STD ERROR	1.805	STD DEV	15.737
VARIANCE	247.651	KURTOSIS	-0.095	SKEWNESS	0.737
RANGE	63.000	MINIMUM	13.000	MAXIMUM	76.000
SUM	2740.000				

VALID OBSERVATIONS - 76 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 #AGGLOMERATIONS

MEAN	1.513	STD ERROR	0.254	STD DEV	2.218
VARIANCE	4.920	KURTOSIS	1.786	SKEWNESS	1.605
RANGE	9.000	MINIMUM	0.0	MAXIMUM	9.000
SUM	115.000				

VALID OBSERVATIONS - 76 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	41.118	STD ERROR	29.752	STD DEV	259.369
VARIANCE	67272.026	KURTOSIS	44.112	SKEWNESS	6.566
RANGE	1944.000	MINIMUM	0.0	MAXIMUM	1944.000
SUM	3125.000				

VALID OBSERVATIONS - 76 MISSING OBSERVATIONS - 0

```
TRANSPACE REQUIRED..      100 BYTES
    1 TRANSFORMATIONS
    0 RECODE VALUES + LAG VARIABLES
   16 IF/COMPUTE OPERATIONS

CPU TIME REQUIRED..    0.28 SECONDS


   15 TASK NAME        CONDESCRIPTIVE WITHOUT FOUR CITY CASES BY SUBFILE
   16 *SELECT IF       (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
   17 RUN SUBFILES     EACH
   18 CONDESCRIPTIVE   V449,V452,V454,V462

***** GIVEN WORKSPACE ALLOWS FOR  350 VARIABLES FOR CONDESCRIPTIVE PROBLEM *****

   19 STATISTICS       ALL
```

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	2636.500	STD ERROR	207.829	STD DEV	974.806
VARIANCE	950246.833	KURTOSIS	-0.738	SKEWNESS	0.510
RANGE	3290.000	MINIMUM	1226.000	MAXIMUM	4516.000
SUM	58047.000				

VALID OBSERVATIONS - 22 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	31.045	STD ERROR	2.915	STD DEV	13.671
VARIANCE	186.903	KURTOSIS	-0.886	SKEWNESS	0.555
RANGE	45.000	MINIMUM	13.000	MAXIMUM	58.000
SUM	683.000				

VALID OBSERVATIONS - 22 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 NAGGLOMERATIONS

MEAN	0.682	STD ERROR	0.202	STD DEV	0.945
VARIANCE	0.894	KURTOSIS	1.552	SKEWNESS	1.464
RANGE	3.000	MINIMUM	0.0	MAXIMUM	3.000
SUM	15.000				

VALID OBSERVATIONS - 22 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	0.0	STD ERROR	0.0	STD DEV	0.0
VARIANCE	0.0	KURTOSIS	0.0	SKEWNESS	0.0
RANGE	0.0	MINIMUM	0.0	MAXIMUM	0.0
SUM	0.0				

VALID OBSERVATIONS - 22 MISSING OBSERVATIONS - 0

should use the official administrative units, in this case the nine districts, as your building blocks; require your subregions to consist of contiguous districts; try to take ecological homogeneity and historical associations into account, but look at a range of variables, including cultural patterns; and finally, if you have reason to believe that districts within a given subregion will show contrast, then organize the subfile so that the figures for the separate districts can be shown. Such region/contrasts within are illustrated in Table 2.5 on page 31 and in most of the tables following it. However, for most purposes, it appeared that only the Southeast (composed of Sbeitla and Jelma) typically showed internal contrast, so only that subregion was divided up in defining the "subfiles" that are listed on page 1 of the output and actually implemented on page 5. Page 6 shows the result for the Western subregion; N=80 for the whole file, and there are 24 cases in the Western subfile. Pages 7-10 show the descriptive statistics for the other subregions and the districts comprising the southeastern subregion.

7. Effect of Selecting Out Special Cases

Earlier on, we noted that the three towns (actually four subdistricts) should be eliminated from this sample because they were different in kind from the rest of the subdistricts. Such deletion is easily accomplished by the *SELECT IF command as shown on page 11. Page 12 recalculates the descriptive statistics for the four variables under study with the new, reduced sample of 76. This output should be compared to its counterpart on page 4. Page 13 then calculates the descriptive statistics for each subfile, deleting the four cases. For example, page 14 shows what happens to the western subregion when the deletions are made. The two subdistricts that comprise the town of Thala are dropped out and the number of subdistricts is reduced to 22. Compare this output with its counterpart on page 6. And in general, study all the rest of the outputs and understand

8. To sum up, one goal of the initial exploration is to make basic decisions on how the rest of the analysis will be organized. Two such decisions pertain to the elimination of special cases, (in this instance, the three towns) and the decision on subregions. Both of these require a lot of background information, many preliminary comparisons of pertinent variables, and some general conception of where the analysis is going. You should take your time and discuss it on all sides, because a mistake at this juncture will be relatively costly to correct later on.

Once you have made these decisions you are ready to list out all the variables. If all is well with these (and you have handled the missing values correctly), you can run CONDESCRIPTIVE and FREQUENCIES, as appropriate. However, these have already been done for this data set, and you can find them in the "Data Book" available in Warren 34.

CONTROL CARDS:

The following represent the cards you will need to obtain a basic description of the on-line data set and Exercise One. The numbers below correspond to those on the output sheets.

(1) Job card *This card identifies you and your account number to the computer.*
You will be using a HASP job card for these programs. Columns 16-22 are for your access code, columns 25-44 are for your name, column 52 is for the job priority code, and columns 78-80 are for the destination of the printout.

(2) /*LIM REGION=19ØK (Note; spaces must appear as shown.)
This card tells the computer to allocate a specific amount of memory for your job. Start the card in Column 1.

(3) // EXEC SPSS
This card tells the computer to go to the disk library and find the catalogued set of procedures you will run (i.e., SPSS).

(4) //FTØ3FØØ1 DD UNIT=SYSDA,DSN=Y8H.RSOC,DISP=SHR (Note: "Ø" is a zero.)
This card tells the system that the data set you are working with resides on a disk. The data set is catalogued under the name Y8H.RSOC and must be specified in exactly that manner or the job won't run. Make sure DISP=SHR. It will allow others to access the data at the same time - thus speeding turn-around time for everyone.

The cards that follow these four are the actual SPSS program which the computer already has. You may ignore these for the time being; however, error messages are contained here, often in the beginning as ABEND=... or in a statement at the end of the computer jargon.

```
   1              16
(5) GET FILE       BIGTUN.FIL
```

This card must be prepared in just this manner; without it your job won't run. This card simply tells the system to locate the SPSS file "RSOC" so that you can work on it. It generates the information in the next printed lines. Note that the format for all cards is keyword in columns 1-15 and text in columns 16-80. Also, the "card" number refers to the number of the command, not the actual number of cards, which corresponds to the separate lines.

(6) LIST CASES CASES=80/VARIABLES=V449,V452,V454,V462
This card tells the computer to list the four variables by case for all 80 cases.

(7) TASK NAME CONDESCRIPTIVE OF POP VARIABLES FOR ALL CASES
This names the following task for easier identification. This card is optional.

(8) CONDESCRIPTIVE V449,V452,V454,V462
(9) STATISTICS ALL
These two cards name the procedure to be run and the statistics to be computed for the list of variables, in this case V449,V452,V454, and V462.

(10) TASK NAME CONDESCRIPTIVE BY SUBFILE

(11) RUN SUBFILES EACH
This card tells the computer to carry out the next procedure separately for each subfile.

(12) CONDESCRIPTIVE V449,V452,V454,V462

(13) STATISTICS ALL

(14) TASK NAME CONDESCRIPTIVE WITHOUT FOUR CITY CASES

(15) *SELECT IF (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
This card tells the computer to select those cases where ID does not equal 102002 or 102002 or 103011 or 207004. The starred SELECT IF indicates that this condition is valid for the next procedure only.

(16) RUN SUBFILES ALL
This card tells the computer that all the subfiles should be grouped together for the following procedures. This command invalidates command number 11 unless it is given again later in the program.

(17) CONDESCRIPTIVE V449,V452,V454,V462

(18) STATISTICS ALL

(19) TASK NAME CONDESCRIPTIVE WITHOUT FOUR CITY CASES BY SUBFILE

(20) *SELECT IF (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))

(21) RUN SUBFILES EACH

(22) CONDESCRIPTIVE V449,V452,V454,V462

(23) STATISTICS ALL

(24) //
This card tells the computer that this is the end of the program.

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	2717.500	STD ERROR	273.295	STD DEV	864.235
VARIANCE	746901.833	KURTOSIS	-0.017	SKEWNESS	0.682
RANGE	2842.000	MINIMUM	1456.000	MAXIMUM	4298.000
SUM	27175.000				

VALID OBSERVATIONS - 10 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	28.400	STD ERROR	3.519	STD DEV	11.128
VARIANCE	123.822	KURTOSIS	4.305	SKEWNESS	1.886
RANGE	38.000	MINIMUM	18.000	MAXIMUM	56.000
SUM	284.000				

VALID OBSERVATIONS - 10 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 AGGLOMERATIONS

MEAN	0.100	STD ERROR	0.100	STD DEV	0.316
VARIANCE	0.100	KURTOSIS	10.000	SKEWNESS	3.162
RANGE	1.000	MINIMUM	0.0	MAXIMUM	1.000
SUM	1.000				

VALID OBSERVATIONS - 10 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	0.0	STD ERROR	0.0	STD DEV	0.0
VARIANCE	0.0	KURTOSIS	0.0	SKEWNESS	0.0
RANGE	0.0	MINIMUM	0.0	MAXIMUM	0.0
SUM	0.0				

VALID OBSERVATIONS - 10 MISSING OBSERVATIONS - 0

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	2584.176	STD ERROR	236.226	STD DEV	973.986
VARIANCE	948648.779	KURTOSIS	0.196	SKEWNESS	0.944
RANGE	3402.000	MINIMUM	1489.000	MAXIMUM	4891.000
SUM	43931.000				

VALID OBSERVATIONS - 17 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	39.118	STD ERROR	3.930	STD DEV	16.205
VARIANCE	262.610	KURTOSIS	-0.119	SKEWNESS	0.753
RANGE	58.000	MINIMUM	18.000	MAXIMUM	76.000
SUM	665.000				

VALID OBSERVATIONS - 17 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 #AGGLOMERATIONS

MEAN	0.353	STD ERROR	0.191	STD DEV	0.786
VARIANCE	0.618	KURTOSIS	8.237	SKEWNESS	2.742
RANGE	3.000	MINIMUM	0.0	MAXIMUM	3.000
SUM	6.000				

VALID OBSERVATIONS - 17 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	114.353	STD ERROR	114.353	STD DEV	471.489
VARIANCE	222302.118	KURTOSIS	17.000	SKEWNESS	4.123
RANGE	1944.000	MINIMUM	0.0	MAXIMUM	1944.000
SUM	1944.000				

VALID OBSERVATIONS - 17 MISSING OBSERVATIONS - 0

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	2853.167	STD ERROR	281.098	STD DEV	1192.599
VARIANCE	**********	KURTOSIS	-0.986	SKEWNESS	0.495
RANGE	3932.000	MINIMUM	1033.000	MAXIMUM	4965.000
SUM	51357.000				

VALID OBSERVATIONS - 18 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	40.611	STD ERROR	3.882	STD DEV	16.468
VARIANCE	271.193	KURTOSIS	0.205	SKEWNESS	0.502
RANGE	61.000	MINIMUM	13.000	MAXIMUM	74.000
SUM	731.000				

VALID OBSERVATIONS - 18 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 AGGLOMERATIONS

MEAN	4.611	STD ERROR	0.549	STD DEV	2.330
VARIANCE	5.428	KURTOSIS	-0.893	SKEWNESS	0.030
RANGE	8.000	MINIMUM	1.000	MAXIMUM	9.000
SUM	83.000				

VALID OBSERVATIONS - 18 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	65.611	STD ERROR	65.611	STD DEV	278.364
VARIANCE	77486.722	KURTOSIS	18.000	SKEWNESS	4.243
RANGE	1181.000	MINIMUM	0.0	MAXIMUM	1181.000
SUM	1181.000				

VALID OBSERVATIONS - 18 MISSING OBSERVATIONS - 0

60

VARIABLE V449 TOTAL POPULATION RESIDENTE

MEAN	3208.333	STD ERROR	352.527	STD DEV	1057.582
VARIANCE	**********	KURTOSIS	-0.997	SKEWNESS	0.520
RANGE	3126.000	MINIMUM	1822.000	MAXIMUM	4948.000
SUM	28875.000				

VALID OBSERVATIONS - 9 MISSING OBSERVATIONS - 0

- -

VARIABLE V452 POP DENSITY

MEAN	41.889	STD ERROR	6.174	STD DEV	18.523
VARIANCE	343.111	KURTOSIS	-0.527	SKEWNESS	0.581
RANGE	56.000	MINIMUM	19.000	MAXIMUM	75.000
SUM	377.000				

VALID OBSERVATIONS - 9 MISSING OBSERVATIONS - 0

- -

VARIABLE V454 #AGGLOMERATIONS

MEAN	1.111	STD ERROR	0.455	STD DEV	1.364
VARIANCE	1.861	KURTOSIS	-1.917	SKEWNESS	0.508
RANGE	3.000	MINIMUM	0.0	MAXIMUM	3.000
SUM	10.000				

VALID OBSERVATIONS - 9 MISSING OBSERVATIONS - 0

- -

VARIABLE V462 COMMUNAL POP RESIDENT

MEAN	0.0	STD ERROR	0.0	STD DEV	0.0
VARIANCE	0.0	KURTOSIS	0.0	SKEWNESS	0.0
RANGE	0.0	MINIMUM	0.0	MAXIMUM	0.0
SUM	0.0				

VALID OBSERVATIONS - 9 MISSING OBSERVATIONS - 0

```
TRANSPACE REQUIRED..      100 BYTES
    1 TRANSFORMATIONS
    0 RECODE VALUES + LAG VARIABLES
   16 IF/COMPUTE OPERATIONS

CPU TIME REQUIRED..    0.43 SECONDS


        20 FINISH          (SPSS GENERATED)

   NORMAL END OF JOB.
      20 CONTROL CARDS WERE PROCESSED.
       0 ERRORS WERE DETECTED.
```

EXERCISE 2: RECODING HISTOGRAMS AND BREAKDOWN ANALYSIS

Problem: What are some ways to display the data so as to benefit your analysis and to provide a preliminary output for potential users?

Brief Answer: Exploit as many of the graphic or breakdown options as are available in the package program. A histogram is one of the conventional ways for displaying certain characteristics of a distribution, and it should be used where possible. Unfortunately, it has a somewhat abstract quality so it may be useful to attach the names of some of the subdistricts to particular cases, such as an extremely low or an extremely high case or perhaps one or two at the mean. A second approach is to provide descriptive statistics for each of a small set of separate regions. If the subregions are well known, the variations from one to another will provide a sense of concreteness to many of the numbers.

Background Reading: SPSS Manual, pps. 40-48; 65-69

Output: A, 1-10; B, 1-5.

Commentary: In almost any working situation, the analyst is under immediate pressure for "results." It is rare that administrators or other potential users comprehend how much time it takes to do even simple analysis, what with checking data, preparing it for the computer, putting up with computer delays, constructing even simple indices, and working out appropriate modes of display. It is possible to relieve some of the pressure by putting out a set of relevant descriptions of some of the simpler variables. This is not lost time in analysis, because so much can be learned from this kind of visual exploration. The key principle is to select variables and modes of display that are at least partially familiar to your users. The examples here, histograms and breakdown analysis based on subregions, meet these requirements, but it won't long before SPSS will offer many other options. Indeed, it is possible to

construct bar graphs, pie charts, etc., already if you have the requisite equipment for drawing the lines. Some of these possibilities may be examined later on.

Steps:

1. Page A1 of the output shows how you make histograms for three familiar variables: population, density, and the number of agglomerations (subcommunities).

The basic data derive from the Tunisian census and are generally familiar. These variables are discussed on pages 44ff of Baseline. Turning to the control cards, you note on line 3 four cases that we have already identified as "special" are dropped. This must be done from now on. Line 4 shows a simple Compute that assigns the name NPOP to V449. Similarly, for NDEN. This procedure corresponds to the general rule of numbering the raw variables and naming the working variables. Working variables usually refer to constructed indices, but in this case the raw data is identical to the working variable, so it gets a name.

2. In order to construct a histogram with continuous variables, such as population or density, it is necessary to make categories, and this is done here with the RECODE command. Note that it is generally better to work with continuous variables than with categories, because information is "lost" when you chop up a continuous variable. But in this case we prefer to lose the information in order to gain visual clarity. Paragraph How many categories should you make? In general, you should make three or more, because otherwise you can't see any distribution, but if the number of categories goes over ten, the picture gets complex and is not much different from a continuous distribution. Of course, when you have a lot of categories, the number of cases in each one is necessarily small. If the number of cases falls below five, reliability is undermined. One last question: How

do you know what the minimum and maximum values are? You already know the answer to that question, having studied the Condescriptives.

3. The first histogram appears on page 4. The analyst has used eight categories here, and the distribution is only slightly skewed to the left (i.e. the cases tend to pile up on the left-hand or low side of the graph.) That is what one would expect in a semi-arid region where the population of many of the subdistricts is low. Indeed, the modal 1 category is 2001-2500 with 23 cases. At the foot of this histogram are the condescriptives for the recoded values. These are useful as a check, but they are misleading except in this particular exercise, so it is preferable to use the condescriptives for the raw values.

The figures for density are shown on page B3 and should be compared to the density of Tunisia as a whole, which is 36 per K^2. The mean density of 36 for this region is considered high for a semi-arid zone, but such judgments imply a whole range of conditions and, of course, vary with such conditions. Probably this histogram would be enhanced by listing the names of several of the least dense and most dense subdistricts.

Page 8 shows the histogram for the number of agglomerations--that is, subcommunities in the subdistrict. Such subcommunities include everything from a compound of houses belonging to an extended family to a small village.

It includes all settlements other than the towns that are listed on page 19 of Baseline. The general geographic distribution of these agglomerations is shown on map 2.8 on page 49, and it is immediately obvious that most of them are in the north. There is more rainfall there, (which permits population concentration.) Otherwise they must be distributed across the countryside in small ravines, near springs, or in a protected area, anywhere that nature offers a small advantage. So the number of agglomerations is an index of the relationship of human groups to the environment. Thus, it is no surprise

to see that 39 subdistricts have no agglomerations at all -- the people are scattered here and there in isolated households -- and the number of subdistricts with many agglomerations declines steadily. Needless to say, this one variable is of great importance in any discussion of service delivery. It is much easier and cheaper to deliver services such as electricity or water to people living in clusters, rather than in scattered households.

4. The second procedure shown in section B of the output is called "breakdown analysis," and the results are the same as those provided by CONDESCRIPTIVE except that you can do breakdown analysis in one step. On page 1 of the output, you simply specify BREAKDOWN and then list the tables you want. In this case, we want total population, population density, and the number of agglomerations broken down by the four subregions. Note that V453 specifies the four subregions. You do not have to use the subfiles for regions that are shown above. That subfile is used for analysis within each region.

Page 2 shows the first breakdown. For the northern region you see that the mean population is 2853.1667, but of course you ignore those last four digits.

The standard deviation and variance are also listed. At the far right note that the northern region has 18 cases--after the one major town has been deleted by the SELECT IF command. The means for the other subregions are also shown, but it is surprising that the more arid southeast has a mean population of 2950. The large town of Sbeitla has been deleted, so this population must derive from the agglomerations that border Sidi Bou Zid. Now, it goes without saying that you would not present this whole table to some administrator. He does not want to see all these numbers. Indeed, at this point only the mean is of much interest, and that should be rounded off. It also goes without saying that you could have used another variable instead of regions for your breakdown, but you wouldn't want to do that at this point, (because your readers are not prepared for it.)

5. How did we decide on the regions? In general, it was a process of trial and error. You read the reports and try to see what subregional boundaries have been used before, you talk to people, you study the maps with respect to ecology and transportation flows, and finally you look at the distributions for the nine districts to see which combinations of these districts are more or less homogeneous. You combine only neighboring districts, and then make assumptions when the values of variables within regions diverges. It is for that reason that the subfiles show the two districts -- Sbeitla and Jelma (which for some reason is spelled Jend.)

6. Page 3 shows the breakdown for population density, and as expected the North has the highest density (41) although it is closely followed by the central region.

The last breakdown deals with a count of structures and is easier to comprehend. Thus, the North has a total of 83 agglomerations, with a mean of about 5. The other regions have markedly few agglomerations.

After this exercise, you will understand those T-shirts that say on the back: "Broken down by age and sex."

```
*****************SYSMSG    ***********RM3.RD1 ***FMT <00>* *****190 LINES ***  6:24.22PM  16 NOV 81  WARREN                SUSAN BENTLEY

B | STEP NAME                    SPSS    |
B | CPU SECONDS                      1.23 |
B | CPU AVGS (CPU*5.00)              6.15 |
B | TOTAL I/O TIME, SECS            20.47 |
B | REGION USED                      148K |
B ***********************************************************************************************************************
B * BEST CLASS FOR JOB- A *|          USED     LIMIT          MIN:SEC                                                    *
B *                       *|                                   USED   LIMIT  $CHARGED          REGION                TOTAL  |
B * AMOUNT SAVED BY       *| LINES.     271     3000                                                                  EXECUTION |
B * RUNNING IN THIS CLASS *| CARDS        0      500   CPU     :01    2:00      $.16   MAX USED   148K              CHARGE  |
B * WOULD BE:        $.02 *| FILING       0      500   I/O     :20    5:00      $.56   LIMIT      160K              ------- |
B ***********************************************************************************************************************
                                                                                                   CHARGE     $.74        $1.46 |
                                                                                                                              *

0 JOB 0763-EM186      FROM RM3.RD1    13:31.27 MONDAY NOVEMBER 16, 1981
0                      //EM186    JOB Y8H$$$$,'SUSAN BENTLEY          ',PRTY=5                          DLVR=WRN
C                      /*LIMITS REGION=160K
C                      // EXEC SPSS
C                      //FT05F001  DD  UNIT=SYSDA,DISP=SHR,DSN=Y8H.RSOC
C                        INPUT STREAM DATASET OF 21 CARDS
C                      //

B *-----------------------------------ACCOUNTING CONTROL RECORD (ACR) DISPLAY-----------------------------------------*
B | Y8H   R S 213 RW                ACCOUNT NO.:  U964F7AA95     STATUS:   ACTIVE       ACCT BALANCE:      $179.58          |
C | DFLT RETPD: SYS          DSACCESS: SYS  DASD+CF PER DAY CHG:   $.13  DASD+CF TOT CHG:      $17.46                   |
C | HASP TOTAL CHG:        $98.02                HASP CHG LIMIT:      $300.00                HASP BALANCE:      $201.98  |
B *-----------------------------------------------------------------------------------------------------------------------*
C //EM186    JOB Y8H$$$$,'SUSAN BENTLEY          ',PRTY=5                                    DLVR=WRN
C // EXEC SPSS
C XXSPSS PROC SPACE=20K,TYPE1=800,TYPE2=2012,TYPE3=TRK,                                      00000010
C XX      MAIN1=250,MAIN2=750,MAIN3=80,MAIN4=2250,                                           00000020
C XX      SEC1=50,SEC2=190,LIB='SYS2.SPSS.LOAD',VERSION=M,                                   00000030
C XX      ARCHIVE='DUMMY,'                                                                   00000040
C ***                                                                                        00000040
C *** CURRENTLY ACCESSING THE INSTALLED VERSION OF SPSS 8.1 (05/20/80)*** 00000050
C *** PROBLEMS SHOULD BE REPORTED TO USER SERVICES AT 607-256-4981.   *** 00000060
C ***                                                                 *** 00000070
C *** PROCEDURE   DEFAULTS  ALLOW   SUFFICIENT   SCRATCH  SPACE   FOR *** 00000080
C *** PROCESSING UP TO  1500 CASES WITH 250 VARIABLES.  OR MORE CASES *** 00000090
C *** WITH  FEWER VARIABLES.   IF  MORE SPACE  IS  NEEDED, USERS  MAY *** 00000100
C *** INCREASE IT ON THE FT01, FT02, &  SORTWK JCL CARDS BY USING THE *** 00000110
C *** MAIN1, MAIN2, MAIN3, SEC1, SEC2, & TYPE3 SUBSTITUTION PARAMETERS*** 00000120
C ***                                                                 *** 00000130
C *** IF 'CLT ARCHIVE', 'SAVE ARCHIVE', OR 'MERGE FILES' IS TO BE USED*** 00000140
C *** THEN CODE ARCHIVE=  ON THE EXEC CARD IN ORDER TO ACTIVATE THE   *** 00000150
C *** NECESSARY SCRATCH SPACE.                                        *** 00000160
C ***                                                                 *** 00000170
C *** FOR THE MAXI VERSION OF SPSS CODE THE FOLLOWING ON THE EXEC CARD*** 00000180
C ***                    // EXEC SPSS,VERSION=M                      *** 00000190
C XXSPSS     EXEC PGM=SPSS&VERSION,PARM=&SPACE                       *** 00000200
C IEF653I SUBSTITUTION JCL - PGM=SPSSM,PARM=20K                          00000210
C XXSTEPLIB  DD DSN=SYS2.LIB,DISP=SHR
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSS.LOAD,DISP=SHR                 00000220
C XXFT01F001 DD UNIT=SYSSC,SPACE=(&TYPE1,(&MAIN1,&SEC1)),DSN=&&LABELS,   00000230
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(800,(250,50)),DSN=&LABELS,
C XX            DISP=(,DELETE)
C XXFT02F001 DD UNIT=SYSSC,SPACE=(&TYPE2,(&MAIN2,&SEC2)),DSN=&&DATASPSS, 00000240
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(2012,(750,190)),DSN=&DATASPSS, 00000250
C XX            DISP=(,DELETE)
C XXFT05F001 DD DDNAME=SYSIN                                             00000260
C XXFT06F001 DD SYSOUT=A                                                 00000270
C XXFT10F001 DD DSN=SYS2.SPSS&VERSION..STAT,DISP=SHR                     00000280
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSSM.STAT,DISP=SHR                00000290
```

A

```
*******SPSS      FT06F001 ************RMS.PRI ***FMT <00>* *****271 LINES ***  6:24.35PM  16 NOV 81  WARREN          SUSAN BENTLEY
SS BATCH SYSTEM                                                                                  11/16/81        PAGE   1
SS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

                          CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM
CER FROM MCGRAW-HILL:  SPSS, 2ND ED. (PRINCIPAL TEXT)        ORDER FROM SPSS INC.:  SPSS STATISTICAL ALGORITHMS
                       SPSS PRIMER (BRIEF INTRO TO SPSS)                            SPSS POCKET GUIDE, RELEASE 8
                       SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)                  KEYWORDS: THE SPSS INC. NEWSLETTER

EFAULT SPACE ALLOCATION..     ALLOWS FOR..     25 TRANSFORMATIONS
ORKSPACE      17920 BYTES                     102 RECODE VALUES + LAG VARIABLES
ANSPACE        2560 BYTES                     412 IF/COMPUTE OPERATIONS

          1 RUN NAME      EXERCISE TWO - GRAPHIC DISPLAY
          2 GET FILE      RSUC


                     FILE RSUC     HAS 409 VARIABLES

                     THE SUBFILES ARE..

                              NO OF
                     NAME     CASES

                     WEST      24
                     SOUT      11
                     CNTR      17
                     NRTH      19
                     JEND       9


PU TIME REQUIRED..    0.57 SECONDS


          3 SELECT IF     (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
          4 COMPUTE       NPOP=V449
          5 COMPUTE       NDEN=V452
          6 RECODE        NPOP (1033 THRU 1489=1) (1510 THRU 1926=2) (2005 THRU 2487=3)
          7                (2510 THRU 2883=4) (3101 THRU 3465=5) (3691 THRU 3912=6)
          8                (4049 THRU 4449=7) (4516 THRU 4965=8)/NDEN (13 THRU 19=1) (20
          9                THRU 29=2) (30 THRU 39=3) (40 THRU 47=4) (50 THRU 59=5) (72 THRU
         10                76=6)/
         11 VAR LABELS     NPOP, RECODED VARIABLES NOW WITH VALUES 1 THRU 8/NDEN, RECODED TO
         12                1 THRU 6
         13 VALUE LABELS   NPOP (1)LESS THAN 1500 (2)BTWN 1500 AND 2000 (3)BTWN 2001 AND
         14                2500 (4)BTWN 2501 AND 3000 (5)BTWN 3001 AND 3500 (6)BTWN 3501 AND
         15                4000 (7)BTWN 4001 AND 4500 (8)BTWN 4501 AND 5000/NDEN (1)BTWN 13
         16                AND 19 (2)BTWN 20 AND 29 (3)BTWN 30 AND 39 (4)BTWN 43 AND 47 (5)
         17                BTWN 50 AND 59 (6)BTWN 72 AND 76
         18 TASK NAME      HISTOGRAM OF POP, DENSITY, AND SUBCOMMUNITIES
         19 FREQUENCIES    GENERAL=NPOP,NDEN,V454
         20 OPTIONS        6
```

GIVEN WORKSPACE ALLOWS FOR 1280 VALUES AND 384 LABELS PER VARIABLE FOR 'FREQUENCIES'

EXERCISE TSC - CROWDING DISPLAY
HISTOGRAM OF FCI, DENSITY, AND SUBCOMMUNITIES 11/16/81 PAGE 3
FILE RSIC (CREATION DATE = 11/06/81)
SUBFILE WEST SOFI CNTR NRTH JEND

NPCP RECODED VARIABLES NOW WITH VALUES 1 THRU

CATEGORY LABEL	CODE	ABSOLUTE FREQ	RELATIVE FREQ (PCT)	ADJUSTED FREQ (PCT)	CUM FREQ (PCT)
LESS THAN 1500	1.	5	6.6	6.6	6.6
BTWN 1500 AND 2000	2.	13	17.1	17.1	23.7
BTWN 2001 AND 2500	3.	23	30.3	30.3	53.9
BTWN 2501 AND 3000	4.	8	10.5	10.5	64.5
BTWN 3001 AND 3500	5.	9	11.8	11.8	76.3
BTWN 3501 AND 4000	6.	6	7.9	7.9	84.2
BTWN 4001 AND 4500	7.	6	7.9	7.9	92.1
BTWN 4501 AND 5000	8.	6	7.9	7.9	100.0
	TOTAL	76	100.0	100.0	

NFCP RECODED VARIABLES NOW WITH VALUES 1 THRU

```
CODE
     I
  1. ****** (      5)
     I  LESS THAN 1500
     I
     I
  2. ************** (     13)
     I  BTWN 1500 AND 2000
     I
     I
  3. ************************ (     23)
     I  BTWN 2001 AND    2500
     I
     I
  4. ********* (      8)
     I  BTWN 2501 AND 3000
     I
     I
  5. ********** (      9)
     I  BTWN 3001 AND 3500
     I
     I
  6. ******* (      6)
     I  BTWN 3501-AND4000
     I
     I
  7. ******* (      6)
     I  BTWN 4001 AND 4500
     I
     I
  8. ******* (      6)
     I  BTWN 4501 AND 5000
     I
     I.........I.........I.........I.........I.........I
     0        10        20        30        40        50
     FREQUENCY
```

AN	3.987	STD ERR	0.231	MEDIAN	3.370
CE	3.000	STD DEV	2.010	VARIANCE	4.040
RTOSIS	-0.662	SKEWNESS	0.595	RANGE	7.000
NIMUM	1.000	MAXIMUM	8.000		

LID CASES 76 MISSING CASES 0

```
FT10F001 DD DSN=SYS2.SPSSAVERSION..STATDISP=SHR
F653I SUBSTITUTION JCL - DSN=SYS2.SPSSH.STAT.DISP=SHR
```

RCISE TWO - GRAPHIC DISPLAY
TOGRAM OF PCP, DENSITY, AND SUBCOMMUNITIES 11/16/81 PAGE 5
F RSCC (CREATION DATE = 11/04/81)
FILE WEST SBEI CNTR NRTH JEND

N RECODED TO1 THRU 6

LSORY LABEL	CODE	ABSOLUTE FREQ	RELATIVE FREQ (PCT)	ADJUSTED FREQ (PCT)	CUM FREQ (PCT)
A 13 AND 19	1.	12	15.8	15.8	15.8
A 20 AND 29	2.	20	26.3	26.3	42.1
A 30 AND 39	3.	18	23.7	23.7	65.8
: 43 AND 47	4.	8	10.5	10.5	76.3
: 50 AND 59	5.	14	18.4	18.4	94.7
: 72 AND 76	6.	4	5.3	5.3	100.0
	TOTAL	76	100.0	100.0	

FILE WEST SBEI CNTR NRTH JEND

N RECODED TO1 THRU 6

```
CODE
     I
  1. ******************************** (    12)
     I  BTWN 13 AND 19
     I
     I
  2. ************************************************** (    20)
     I  BTWN 20 AND 29
     I
     I
  3. ******************************************** (    18)
     I  BTWN 30 AND 39
     I
     I
  4. ********************* (     8)
     I  BTWN 43 AND 47
     I
     I
  5. ********************************** (    14)
     I  BTWN 50 AND 59
     I
     I
  6. ********** (     4)
     I  BTWN 72 AND 76
     I
     I.........I.........I.........I.........I.........I
     0         4         8        12        16        20
     FREQUENCY
```

	3.053	STD ERR	0.172	MEDIAN	2.833
	2.000	STD DEV	1.496	VARIANCE	2.237
OSIS	-0.967	SKEWNESS	0.350	RANGE	5.000
MUM	1.000	MAXIMUM	6.000		

D CASES 76 MISSING CASES 0

6631 SUBSTITUTION JCL - DSN=SYS2.SPSSH.STAT.DISP=SHR

AGGLOMERATIONS

SORY LABEL	CODE	ABSOLUTE FREQ	RELATIVE FREQ (PCT)	ADJUSTED FREQ (PCT)	CUM FREQ (PCT)
	0.	39	51.3	51.3	51.3
	1.	13	17.1	17.1	68.4
	2.	5	6.6	6.6	75.0
	3.	7	9.2	9.2	84.2
	4.	3	3.9	3.9	88.2
	5.	2	2.6	2.6	90.8
	6.	2	2.6	2.6	93.4
	7.	4	5.3	5.3	98.7
	9.	1	1.3	1.3	100.0
	TOTAL	76	100.0	100.0	

V454 AGGLOMERATIONS

```
CODE
     I
  0. ************************************** (    39)
     I
     I
     I
  1. ************** (    13)
     I
     I
     I
  2. ****** (     5)
     I
     I
     I
  3. ******** (     7)
     I
     I
     I
  4. **** (     3)
     I
     I
     I
  5. *** (     2)
     I
     I
     I
  6. *** (     2)
     I
     I
     I
  7. ***** (     4)
     I
     I
     I
  9. ** (     1)
     I
     I
     I..........I.........I.........I.........I.........I
     0         10        20        30        40        50
     FREQUENCY
```


C IEF653I SUBSTITUTION ...ICL - DSN=SYS2.SPSSH.STAT.DISP=SHR...

EXERCISE TWO - GRAPHIC DISPLAY
HISTOGRAM OF PCP, DENSITY, AND SUBCOMMUNITIES 11/16/81 PAGE 9
FILE RSOC (CREATION DATE = 11/04/81)
SUBFILE WEST SBFI CNTR NRTH JEAD

MEAN	1.513	STD ERR	0.254	MEDIAN	0.474
MODE	0.0	STD DEV	2.210	VARIANCE	4.520
KURTOSIS	1.786	SKEWNESS	1.605	RANGE	9.000
MINIMUM	0.0	MAXIMUM	9.000		

VALID CASES 76 MISSING CASES 0

53

```
TRANSPACE REQUIRED..      400 BYTES
     4 TRANSFORMATIONS
    16 RECODE VALUES + LAG VARIABLES
    20 IF/COMPUTE OPERATIONS

CPU TIME REQUIRED..     0.33 SECONDS

          22 FINISH ------- (SPSS GENERATED)

NORMAL END OF JOB.
    22 CONTROL CARDS WERE PROCESSED.
     0 ERRORS WERE DETECTED.
```

```
C XX            DISP=(,DELETE,...
C XXFT05F001 DD DDNAME=SYSIN
C XXFT06F001 DD SYSOUT=A
C XXFT10F001 DD DSN=SYS2.SPSS&VERSION..STAT,DISP=SHR                00000280
                                                                    00000290
```

```
**********************SYSMSG  ************RM3.PR1 ***FMT <00>*******198 LINES*** 11:12.13PM  16 DEC 81  WARREN                    SUSAN BENTLEY
B | STEP NAME                    SPSS  |
B | CPU SECONDS                   1.07 |
B | CPU AUTS (CPU*5.00)           5.35 |
B | TOTAL I/O TIME, SECS         19.42 |
B | REGION USED                  148K |
B ***************************************************************************************************************
B * BEST CLASS FOR JOB- A *|          USED    LIMIT          MIN:SEC                                  REGION          TOTAL  |
B *                       *|                                   USED   LIMIT  $CHARGED                                EXECUTION |
B * AMOUNT SAVED BY       *| LINES      98     3000                                                  MAX USED   140K    CHARGE  |
B * RUNNING IN THIS CLASS *| CARDS       0      500   CPU      :01    2:00      $.10                 LIMIT      160K            |
B * WOULD BE:        $.02 *| FILING      0      500   I/O      :19    5:00      $.38                 CHARGE     $.49      $.97  |
B ***************************************************************************************************************

D JOB 0331-EN142     FROM RM3.RD1   12:39.22 WEDNESDAY DECEMBER 16, 1981
D                    //EN142   JOB Y8H$$$$,'SUSAN BENTLEY        ',PRTY=4                    DLVR=WRN
C                    /*LIMITS REGION=160K
C                    // EXEC SPSS
C                    //FT03F001  DD  UNIT=SYSDA,DISP=SHR,DSN=Y8H.RSOC
C                      INPUT STREAM DATASET OF 3 CARDS
C                    //

B *-----------------------------------ACCOUNTING CONTROL RECORD (ACR) DISPLAY-----------------------------------*
B | Y8H   R S 213 NW              ACCOUNT NO.:  U964F7AA99    STATUS:   ACTIVE      ACCT BALANCE:     $128.28      |
C | DFLT RETPD: SYS        DSACCESS: SYS  DASD+CF PER DAY CHG:  $.13  DASD+CF TOT CHG:      $26.29              |
C | HASP TOTAL CHG:     $138.89                HASP CHG LIMIT:     $300.00                HASP BALANCE:     $161.11  |
B *-------------------------------------------------------------------------------------------------------------*
C //EN142    JOB Y8H$$$$,'SUSAN BENTLEY        ',PRTY=4                                  DLVR=WRN
C // EXEC SPSS
C XXSPSS PROC SPACE=20K,TYPE1=800,TYPE2=2012,TYPE3=TRK,                                  00000010
C XX       MAIN1=250,MAIN2=750,MAIN3=80,MAIN4=2250,                                      00000020
C XX       SEC1=50,SEC2=190,LIB='SYS2.SPSS.LOAD',VERSION=H,                              00000030
C XX       ARCHIVE='DUMMY,'                                                              00000040
C ***                                                                                    *** 00000050
C *** CURRENTLY ACCESSING THE INSTALLED VERSION OF SPSS 8.1 (05/20/80)*** 00000060
C *** PROBLEMS SHOULD BE REPORTED TO USER SERVICES AT 607-256-4981.   *** 00000070
C ***                                                                                    *** 00000080
C *** PROCEDURE   DEFAULTS  ALLOW   SUFFICIENT   SCRATCH  SPACE   FOR *** 00000090
C *** PROCESSING UP TO  1500 CASES WITH 250 VARIABLES,  OR MORE CASES *** 00000100
C *** WITH  FEWER VARIABLES.   IF  MORE SPACE  IS  NEEDED, USERS  MAY *** 00000110
C *** INCREASE IT ON THE FT01, FT02, &  SORTWK JCL CARDS BY USING THE *** 00000120
C *** MAIN1, MAIN2, MAIN3, SEC1, SEC2, & TYPE3 SUBSTITUTION PARAMETERS*** 00000130
C ***                                                                                    *** 00000140
C *** IF 'GET ARCHIVE', 'SAVE ARCHIVE', OR 'MERGE FILES' IS TO BE USED*** 00000150
C *** THEN CODE ARCHIVE=  ON THE EXEC CARD IN ORDER TO ACTIVATE THE   *** 00000160
C *** NECESSARY SCRATCH SPACE.                                                           *** 00000170
C ***                                                                                    *** 00000180
C *** FOR THE MAXI VERSION OF SPSS CODE THE FOLLOWING ON THE EXEC CARD*** 00000190
C ***                    // EXEC SPSS,VERSION=M                                          *** 00000200
C XXSPSS     EXEC PGM=SPSS&VERSION,PARM=&SPACE                                           00000210
C IEF653I SUBSTITUTION JCL - PGM=SPSSH,PARM=20K
C XXSTEPLIB  DD DSN=&LIB,DISP=SHR                                                        00000220
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSS.LOAD,DISP=SHR
C XXFT01F001 DD UNIT=SYSSC,SPACE=(&TYPE1,(&MAIN1,&SEC1)),DSN=&&LABELS,                   00000230
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(800,(250,50)),DSN=&LABELS,
C XX            DISP=(,DELETE)                                                           00000240
C XXFT02F001 DD UNIT=SYSSC,SPACE=(&TYPE2,(&MAIN2,&SEC2)),DSN=&&DATASPSS,                 00000250
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(2012,(750,190)),DSN=&DATASPSS,
C XX            DISP=(,DELETE)                                                           00000260
C XXFT05F001 DD DDNAME=SYSIN                                                             00000270
C XXFT06F001 DD SYSOUT=A                                                                 00000280
C XXFT10F001 DD DSN=SYS2.SPSS&VERSION,,STAT,DISP=SHR                                     00000290
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSSH.STAT,DISP=SHR
```

```
**********SPSS     FT06F001 ************RM3,PR1 ***FMT <00>* ******98 LINES*** 11:12.28PM  16 DEC 81  WARREN          SUSAN BENTLEY
SPSS BATCH SYSTEM                                                                                  12/16/81          PAGE    1

SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

                    CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM
ORDER FROM MCGRAW-HILL:  SPSS, 2ND ED. (PRINCIPAL TEXT)        ORDER FROM SPSS INC.:  SPSS STATISTICAL ALGORITHMS
                         SPSS PRIMER (BRIEF INTRO TO SPSS)                             SPSS POCKET GUIDE, RELEASE 8
                         SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)                   KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION..      ALLOWS FOR..    25 TRANSFORMATIONS
WORKSPACE     17920 BYTES                     102 RECODE VALUES + LAG VARIABLES
TRANSPACE      2560 BYTES                     412 IF/COMPUTE OPERATIONS

                1 GET FILE      RSOC


                          FILE RSOC     HAS 409 VARIABLES

                          THE SUBFILES ARE..

                                   NO OF
                          NAME     CASES

                          WEST       24
                          SDEI       11
                          CNTR       17
                          NRTH       19
                          JEND        5


CPU TIME REQUIRED..     0.51 SECONDS


                2 SELECT IF       (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
                3 BREAKDOWN       TABLES=V445,V452,V454 BY V453

***** GIVEN WORKSPACE ALLOWS FOR    639 CELLS AND     1 DIMENSIONS FOR SUBPROGRAM BREAKDOWN *****
```

60

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

DESCRIPTION OF SUBPOPULATIONS

CRITERION VARIABLE V449 TOTAL POPULATION RESIDENTE
BROKEN DOWN BY V453 REGION

VARIABLE	CODE	VALUE LABEL	SUM	MEAN	STD DEV	VARIANCE	N
FOR ENTIRE POPULATION			209385.0000	2755.0658	1018.5889	1037523.3956	(76)
V453	1.	NORTH	51357.0000	2853.1667	1192.5986	1422291.3235	(18)
V453	2.	WEST	58047.0000	2638.5000	974.8060	950246.8333	(22)
V453	3.	CENTRAL	43931.0000	2584.1765	973.9860	948648.7794	(17)
V453	4.	SOUTH	56050.0000	2950.0000	966.4121	933952.3333	(19)

TOTAL CASES = 76

6/

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

DESCRIPTION OF SUBPOPULATIONS

CRITERION VARIABLE V452 POP DENSITY
BROKEN DOWN BY V453 REGION

VARIABLE	CODE	VALUE LABEL	SUM	MEAN	STD DEV	VARIANCE	N
FOR ENTIRE POPULATION			2740.0000	36.0526	15.7369	247.6505	(76)
V453	1.	NORTH	731.0000	40.6111	16.4679	271.1928	(18)
V453	2.	WEST	683.0000	31.0455	13.6712	186.9026	(22)
V453	3.	CENTRAL	665.0000	39.1176	16.2053	262.6103	(17)
V453	4.	SOUTH	661.0000	34.7895	16.1953	262.2865	(19)

TOTAL CASES = 76

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBCI CNTR NRTH JEND

DESCRIPTION OF SUBPOPULATIONS

CRITERION VARIABLE V454 #AGGLOMERATIONS
BROKEN DOWN BY V453 REGION

VARIABLE	CODE	VALUE LABEL	SUM	MEAN	STD DEV	VARIANCE	N
FOR ENTIRE POPULATION			115.0000	1.5132	2.2181	4.9198	(76)
V453	1.	NORTH	83.0000	4.6111	2.3298	5.4281	(18)
V453	2.	WEST	15.0000	0.6818	0.9455	0.8939	(22)
V453	3.	CENTRAL	6.0000	0.3529	0.7859	0.6176	(17)
V453	4.	SOUTH	11.0000	0.5789	1.0706	1.1462	(19)

TOTAL CASES = 76

63

EXERCISE 3: INSPECTING THE RAW DATA AND PERCENTAGES

Problem: How does one handle the raw data that is to be used for a measure of poverty?

Short answer: List it out and inspect it. Assume that there will be more errors in this informant-produced data than in the census data in Exercises 1 and 2. Eliminate unuseable or irrelevant variables. Then follow an appropriate procedure for constructing an index. In this case, that procedure involves: (a) percentaging, (b) test for clustering, (c) computation of "z" scores, and (d) combining the "z" scores into a composite score. However, this exercise deals only with the first step.

Output: Pages A 1-6

Background: Although a good baseline should include a measure of poverty, such an index is particularly difficult to construct on the basis of informant data. However, it had to be done because there was no possibility of administering a sample survey in this district; and the census data, although it contained variables dealing with housing, was not available. That is, it was not published for the subdistrict level, and a request for a special run was ignored. To make things more difficult, AID, the sponsor of this project, wanted a measure of poverty of the poorest segment of the population. Asking these Tunisian farmers and herders how much money they earn during a year made no sense, because much of their income was in kind. So we attempted to construct an index of housing characteristics and household possessions by asking the informants for information. The details of this index and some of the problems of field work are given in Baseline, p. 55 ff. In brief, we decided to ask about the household characteristics and possessions of the farmers who had no irrigation but depended on rain. These are referred to as the "dryland farmers."

Steps

1. Print out the relevant variables, which in this case are variables 211 and 215-224. In the English version of the questionnaire, these items read as follows:

V211 How many dry farmers' houses are there?

V215 How many dry farmers' houses have glass windows?

V216 A commercially made door

V217 A gas stove

V218 A sewing machine

V219 A cookstove in the house

V220 Two stories

V221 A cement or paved floor

V222 A tile roof

V223 A separate oven

V224 Latrines outside the house

2. In scanning this listing note first the six-digit serial numbers at the left. In Exercises 1 and 2 we used the serial numbers that the computer generated, 1-80, but those cannot be relied on in future exercises, because we are planning to delete four cases. So it is important to use the official serial numbers, which have three parts. The first digit refers to the "state" (gouvernorat) in which the subdistrict is located. The next two digits refer to the district location, and they go from 01 to 09. And finally, the last three digits refer to the particular subdistrict within these two larger administrative units. Consult the list of serial numbers in the first pages of these exercises for the names of these places.

You may ask, why use such a cumbersome numbering system? The main reason is that it can be expanded for the whole of Central Tunisia, consisting of five states, about 15 districts, and 200 subdistricts. Another reason for using this system is that at the outset of research we literally did not know the names or locations of the subdistricts, and neither did the Tunisians. This might surprise you, but it is a frequent problem. Even in Puerto Rico the official maps do not match all the names and locations that one finds on the ground. Therefore, one needs an open-ended system of serial numbers so that "newly discovered" subdistricts can be added in the proper place.

The second column, V211, refers to the informant's estimate of the number of dryland farmers in his subdistrict. That number will serve as the denominator for the percentages. The other numbers refer to the informant's estimates of the number of particular attributes, as listed above. You may well wonder about the accuracy of these estimates, and I can assure you that there was a great deal of debate about how to get

this information, or, for that matter, whether to attempt it at all. On the second question, we felt we had no choice. Any baseline that did not produce some estimate of poverty was not an adequate baseline, at least according to prevalent conceptions of rural development. The only question then was how to get the information. We were persuaded that these local leaders could provide actual numbers, and we acted on this advice. The alternatives to asking for raw numbers were to ask for percentages (but that demanded even more quantitative precision on the part of the informant) or to use verbal categories such as none, a few, some, etc. We might have indicated percentage ranges for each of these verbal categories. Each alternative has advantages and disadvantages; for better or worse, the numbers won out. However, one still does not have to take these numbers literally. As the analysis proceeds, it will be apparent that they are being used as rough approximations. People who have not had methodological training often believe that quantitative analysts take numbers like these literally, but the truth is just the opposite. It is usually the administrators who get fixated on particular numbers, not realizing the measurement error involved and the general limits to interpretation.

3. Inspect the list for "problems":

<u>None or too few responses</u>. On p. 4, variable 220 (two stories) and variable 224 (latrines outside the house) have no positive responses. These attributes are not present in any of the houses of dryland farmers. Consequently these items are useless in analyzing the characteristics of dryland farmer poverty. (Such items may still be relevant to the assessment of the houses of the irrigated farmers.)

<u>No information for the subdistrict</u>. Notice that subdistricts 102002,

105008 and 207004 have no information at all and were coded as such with 9's. What happened here? Did the informant get tired, or did the interviewer neglect to ask these questions? One clue to what happened is the fact that two of these places are district seats. Perhaps the informant felt he could not estimate the number of houses with given attributes in places that were so densely populated. Note, however, that we will eliminate two of the three cases as "towns." The other will be coded as missing data (using -1) and dropped from this part of the analysis. That leaves 103009, where data on dryland farmer population is missing. That too generates a row of "-1's," but in this case a special effort should have been made to estimate this figure, so that all the rest of the data could be saved.

Extreme values or unusual distributions. V221 (cement or paved floor) poses another problem. The number of answers is so few as to be unusable, but, more important, three of the eight answers are extreme values, specifically 100, 200 and 300. Even if the informant is correct in reporting so many houses with paved floors in his subdistrict, these extreme values will distort subsequent statistical analysis.

Irrelevant variables. V222 (tiled roofs) was included here as an indicator of French presence during the colonial period and was not expected to serve as an indicator. Even if a Tunisian farmer happens to occupy a house with a tiled roof, it is unlikely that such an indicator correctly reflects his present level of living. So this item was dropped from this analysis but may be relevant to an analysis of French influence.

4. Compute percentages as shown in the control cards and the output on page 5. However, before proceeding to this step it is necessary to study the figures on the population of dryland farmers (V211) quite carefully. Are they reasonable? Is it really possible that there are only 8 dryland farmers

in subdistrict 101002? Or 0 in 102002? In fact, we already know the reason why there is a 0 in this particular case. What about 3,000 in 104002 and 3,200 in 309001?

We can check these numbers in two ways: an external source or internal consistency. As it happened, there was no external source because the Tunisian census is not refined in this way. That leaves internal consistency, such as checking to see whether the number of dry-land farmers is greater than the estimated total population or that the percentages for the housing attributes come to more than 100 percent. Some corrections can be made on the basis of this type of analysis, but the truth is that we are dealing with a great deal of measurement error here. One way to handle such numbers is to recode them to 100 percent.

There are a number of considerations. Even if particular numbers are incorrect, it may still be possible to use them as general indicators of level in broad characterizations or in correlations. They may be accurate enough, for example, to show which subregion is high and which one is low. As for correlations, if one finds a strong relationship between, say, the proportion of houses with windows and the level of urbanization, then it is likely to be a generally correct correlation, because it is unlikely that the reporting errors were systematically biased. If they were random, sometimes high and sometimes low, they would tend to cancel each other out and reduce any correlation coefficient. If the numbers are systematically biased, that can often be discovered in the course of analysis.

5. Inspect the computed percentages, as shown on pages 5 and 6. Note first that the missing data code has taken care of some of our problems, and the deletion of the three towns will drop one more case. The next thing to

note is that the percentages for ovens are generally quite high, raising the question of whether this item is usable. If almost everybody has an oven of some sort, then the item will not discriminate the rich from the poor houses, and it probably will not correlate very well with the other attributes. The same points hold for DSEW, the proportion of dryland farmers' houses with sewing machines. However, these questions cannot be settled yet. They are taken up in a subsequent exercise.

For the moment we are looking for discrepancies, and clearly there is one in subdistrict 101002. According to these figures, none of these houses have windows, stoves, or sewing machines; but 100 percent have doors and ovens. Not very likely. Another suspicious case is 104003. Indeed, the one below it has approximately the same pattern.

We could decide to delete a lot of these cases, just as we have the three towns, but then one faces the painful prospect of working with a reduced sample. It is all the more painful if the data are otherwise adequate for these cases. One might be tempted, at this point, to recode an odd 100 percent if the other three or four attributes are all 0. In certain cases, this kind of arbitrary procedure is justified, and one of the nice features of SPSS is that one can keep a record of all of these changes. However, it is, in general, a bad practice, because it quickly gets out of control and you don't know where you stand. The other procedure is to rely on the corrective power of the composite indicators. That is, we could make an indicator of housing characteristics by adding up three, four, or five of these attributes, and that would have the effect of diluting any extreme values (like 100 percent) if all the other values were low (0 percent). These questions will be taken up in a later exercise.

```
*****************SYSMSG    ************RM3.PR1 ***FMT <00>* *****201 LINES ***  5:57.54PM  17 NOV 81  WARREN                SUSAN BENTLEY
B | STEP NAME                    SPSS    |
B | CPU SECONDS                     1.63 |
B | CPU AU'S (CPU*5.00)             8.15 |
B | TOTAL I/O TIME, SECS           23.82 |
B | REGION USED                     152K |
B ***************************************************************************************************************
B * BEST CLASS FOR JOB- A *|          USED     LIMIT              MIN:SEC                                          *
B *                       *|                                     USED   LIMIT  $CHARGED          REGION            TOTAL  |
B * AMOUNT SAVED BY       *| LINES     229      3000                                                                 |
B * RUNNING IN THIS CLASS *| CARDS       0       500   CPU      :02    2:00       $.22   MAX USED     152K   EXECUTION |
B * WOULD BE:        $.02 *| FILING      0       500   I/O      :24    5:00       $.65   LIMIT        160K   CHARGE  |
B ***************************************************************************************************************
                                                                                                                      *
0 JOB 6221-EM190      FROM RM3.RD1   12:06.28 TUESDAY NOVEMBER 17, 1981                                   $1.76 |
0                      //EM190    JOB Y8H$$$$,'SUSAN BENTLEY          ',PRTY=5                                  *
C                      /*LIM REGION=160K                                                    DLVR=WRN
C                      // EXEC SPSS
C                      //FT03F001  DD  UNIT=SYSDA,DISP=SHR,DSN=Y8H.RSOC
C
C                        INPUT STREAM DATASET OF 15 CARDS
C                      //
```

Page A

```
B *-----------------------------------ACCOUNTING CONTROL RECORD (ACR) DISPLAY-----------------------------------*
B | Y8H   R S 213 NW                ACCOUNT NO.:  U964F7AA99    STATUS:   ACTIVE      ACCT BALANCE:      $174.64   |
C | DFLT RETPD: SYS        DSACCESS: SYS  DASD+CF PER DAY CHG:   $.13  DASD+CF TOT CHG:      $17.78              |
C | HASP TOTAL CHG:      $102.64                HASP CHG LIMIT:     $300.00                HASP BALANCE:     $197.36 |
B *-----------------------------------------------------------------------------------------------------------------*
C //EM190    JOB Y8H$$$$,'SUSAN BENTLEY          ',PRTY=5                                  DLVR=WRN
C // EXEC SPSS
C XXSPSS PROC SPACE=20K,TYPE1=800,TYPE2=2012,TYPE3=TRK,                                    00000010
C XX       MAIN1=250,MAIN2=750,MAIN3=80,MAIN4=2250,                                        00000020
C XX       SEC1=50,SEC2=190,LIB='SYS2.SPSS.LOAD',VERSION=H,                                00000030
C XX       ARCHIVE='DUMMY,'                                                                00000040
C ***                                                                                      00000040
C *** CURRENTLY ACCESSING THE INSTALLED VERSION OF SPSS 8.1 (05/20/80)*** 00000050
C *** PROBLEMS SHOULD BE REPORTED TO USER SERVICES AT 607-256-4981.   *** 00000060
C ***                                                                  *** 00000070
C *** PROCEDURE  DEFAULTS  ALLOW  SUFFICIENT  SCRATCH  SPACE  FOR      *** 00000080
C *** PROCESSING UP TO  1500 CASES WITH 250 VARIABLES,  OR MORE CASES  *** 00000090
C *** WITH  FEWER VARIABLES.  IF  MORE SPACE  IS  NEEDED, USERS  MAY   *** 00000100
C *** INCREASE IT ON THE FT01, FT02, &  SORTWK JCL CARDS BY USING THE  *** 00000110
C *** MAIN1, MAIN2, MAIN3, SEC1, SEC2, & TYPE3 SUBSTITUTION PARAMETERS*** 00000120
C ***                                                                  *** 00000130
C *** IF 'GET ARCHIVE', 'SAVE ARCHIVE', OR 'MERGE FILES' IS TO BE USED*** 00000140
C *** THEN CODE ARCHIVE=  ON THE EXEC CARD IN ORDER TO ACTIVATE THE   *** 00000150
C *** NECESSARY SCRATCH SPACE.                                         *** 00000160
C ***                                                                  *** 00000170
C *** FOR THE MAXI VERSION OF SPSS CODE THE FOLLOWING ON THE EXEC CARD*** 00000180
C ***                    // EXEC SPSS,VERSION=M                       *** 00000190
C XXSPSS     EXEC PGM=SPSS&VERSION,PARM=&SPACE                         *** 00000200
C IEF653I SUBSTITUTION JCL - PGM=SPSSH,PARM=20K                                            00000210
C XXSTEPLIB  DD DSN=&LIB,DISP=SHR                                                          00000220
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSS.LOAD,DISP=SHR
C XXFT01F001 DD UNIT=SYSSC,SPACE=(&TYPE1,(&MAIN1,&SEC1)),DSN=&&LABELS,                     00000230
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(800,(250,50)),DSN=&LABELS,
C XX            DISP=(,DELETE)                                                             00000240
C XXFT02F001 DD UNIT=SYSSC,SPACE=(&TYPE2,(&MAIN2,&SEC2)),DSN=&&DATASPSS,                   00000250
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(2012,(750,190)),DSN=&DATASPSS,
C XX            DISP=(,DELETE)                                                             00000260
C XXFT05F001 DD DDNAME=SYSIN                                                               00000270
C XXFT06F001 DD SYSOUT=A                                                                   00000280
```

```
SPSS for DECSYSTEM-20, Version M, Release 9.1, May 1, 1982

                    Current Documentation for the SPSS Batch System
Order from McGraw Hill:  SPSS, 2nd Ed. (principal text)      Order from SPSS Inc.:  SPSS Statistical Algorithms
                         SPSS Primer (brief intro to SPSS)                         KEYWORDS: The SPSS Inc. Newsletter
                         SPSS Update 7-9 (use w/SPSS,2nd for Rel. 7, 8, 9)
                         SPSS Introductory Guide: Basic Statistics and Operations
                         SPSS Primer (Brief Intro to SPSS)

Default SPACE allocation:     Allows for:          98 Transformations
WORKSPACE      17920 words                        394 RECODE values + LAG variables
TRANSPACE       7500 words                       1576 IF/COMPUTE operations

                  1 RUN NAME       EXERCISE THREE
                  2 GET FILE       BIGTUN.FIL

WARNING: Expected file BIGTUN   found file BIGTUNIS.


File BIGTUNIS has  497 variables

The subfiles are:

          No. of
Name      Cases

WEST         24
SBEI         11
CNTR         17
NRTH         19
JEND          9


CPU time required:     0.79 seconds


                  3 COMPUTE        DWIND=(V215/V211)*100
                  4 COMPUTE        DDOOR=(V216/V211)*100
                  5 COMPUTE        DSTOVE=(V217/V211)*100
                  6 COMPUTE        DSEW=(V218/V211)*100
                  7 COMPUTE        DOVEN=(V223/V211)*100
                  8 VAR LABELS     DWIND % DRYLAND FARMHOUSES WITH WINDOWS/DDOOR % DRYLAND
                  9                FARMS WITH DOORS/DSTOVE % DRYLAND FARMHOUSES WITH STOVES
                 10                DSEW % DRYLAND FARMHOUSES WITH SEWING MACHINES/DOVEN
                 11                % DRYLAND FARMHOUSES WITH OVENS
                 12 ASSIGN MISSING DWIND TO DOVEN(-1)
                 13 TASK NAME      EXAMINING THE DATA
                 14 LIST CASES     CASES=80/VARIABLES=ID V211 V215 TO V222
                 15 CONDESCRIPTIVE V215

***** Given WORKSPACE allows for 1378 variables for CONDESCRIPTIVE problem *****
```

File BIGTUNIS (Creation date = 02/04/83)
Subfile WEST SELI CNTR NRTH UFND

Case	ID	V211	V215	V216	V217	V218	V219	V220	V221	V222
1	101001.	800.	144.	400.	2.	0.	0.	0.	25.	0.
2	101002.	8.	0.	8.	0.	0.	0.	0.	0.	0.
3	101003.	325.	10.	25.	50.	6.	0.	0.	0.	1.
4	101004.	700.	300.	300.	10.	8.	0.	0.	100.	0.
5	101005.	996.	80.	400.	200.	2.	0.	0.	0.	10.
6	101006.	450.	60.	60.	15.	10.	0.	0.	0.	0.
7	101007.	10.	10.	10.	2.	0.	0.	0.	0.	0.
8	101008.	300.	70.	210.	0.	4.	1.	0.	0.	1.
9	101009.	250.	200.	250.	70.	4.	5.	0.	0.	0.
10	102001.	62.	0.	0.	0.	0.	0.	0.	0.	0.
11	102002.	0.	999.	999.	999.	999.	999.	999.	999.	999.
12	102003.	350.	200.	200.	20.	5.	10.	0.	10.	100.
13	102004.	600.	200.	200.	40.	15.	0.	0.	0.	3.
14	102005.	320.	60.	250.	80.	4.	4.	0.	0.	40.
15	102006.	319.	150.	200.	30.	0.	0.	0.	0.	0.
16	102007.	350.	100.	70.	50.	3.	3.	0.	0.	6.
17	102008.	250.	10.	10.	0.	0.	0.	0.	0.	0.
18	102009.	472.	50.	150.	60.	0.	5.	0.	0.	5.
19	102010.	110.	5.	5.	0.	0.	2.	0.	0.	1.
20	102011.	150.	100.	100.	0.	3.	0.	0.	5.	10.
21	102012.	484.	50.	50.	70.	0.	0.	0.	0.	10.
22	102013.	500.	75.	75.	60.	3.	3.	0.	0.	4.
23	102014.	200.	50.	50.	0.	0.	0.	0.	0.	0.
24	102015.	300.	18.	18.	60.	5.	3.	0.	0.	14.
25	103001.	890.	60.	60.	30.	1.	0.	0.	0.	0.
26	103002.	549.	186.	186.	10.	4.	0.	0.	0.	1.
27	103003.	360.	30.	120.	5.	0.	0.	0.	0.	0.
28	103004.	500.	60.	60.	5.	0.	0.	0.	0.	1.
29	103005.	470.	200.	470.	5.	3.	0.	0.	0.	0.
30	103006.	500.	40.	40.	10.	1.	2.	0.	0.	0.
31	103007.	600.	350.	350.	4.	0.	0.	0.	0.	0.
32	103008.	500.	50.	50.	20.	2.	0.	0.	0.	0.
33	103009.	9999.	5.	210.	5.	2.	0.	0.	0.	0.
34	103010.	444.	110.	110.	2.	0.	0.	0.	0.	0.
35	103011.	400.	250.	300.	50.	3.	0.	0.	0.	0.
36	104001.	400.	30.	30.	10.	3.	0.	0.	0.	1.
37	104002.	3000.	2.	0.	3.	0.	0.	0.	0.	0.
38	104003.	500.	0.	500.	20.	0.	0.	0.	0.	0.
39	104004.	200.	0.	200.	10.	0.	0.	0.	0.	0.
40	104005.	500.	5.	400.	15.	0.	0.	0.	0.	0.
41	104006.	120.	120.	120.	110.	6.	0.	0.	0.	0.
42	104007.	593.	5.	300.	5.	2.	0.	0.	0.	0.
43	105001.	350.	10.	100.	10.	0.	0.	0.	0.	30.
44	105002.	200.	100.	100.	30.	1.	1.	0.	0.	1.
45	105003.	50.	15.	30.	8.	2.	0.	0.	1.	8.
46	105004.	33.	33.	33.	7.	1.	2.	0.	0.	0.

47	105005.	500.	50.	50.	15.	3.	0.	0.	0.	0.
48	105006.	300.	50.	200.	0.	0.	0.	0.	0.	0.
49	105007.	300.	0.	300.	0.	0.	0.	0.	0.	0.
50	105008.	0.	999.	999.	999.	999.	0.	0.	0.	1.
51	105009.	220.	0.	220.	20.	2.	999.	999.	999.	999.
52	105010.	300.	40.	200.	0.	0.	0.	0.	0.	0.
53	206001.	280.	200.	200.	180.	16.	0.	0.	0.	0.
54	206002.	1400.	1.	90.	0.	0.	0.	0.	0.	0.
55	206003.	157.	20.	20.	10.	1.	0.	0.	0.	0.
56	206004.	475.	0.	47.	0.	0.	0.	0.	0.	5.
57	206005.	400.	0.	400.	0.	0.	0.	0.	0.	0.
58	206006.	450.	100.	100.	0.	0.	0.	0.	0.	1.
59	207001.	80.	80.	80.	5.	0.	0.	0.	0.	0.
60	207002.	1100.	10.	970.	15.	12.	6.	0.	0.	40.
61	207003.	150.	35.	35.	5.	0.	0.	0.	0.	15.
62	207004.	0.	999.	999.	999.	999.	4.	0.	10.	999.
63	207005.	400.	0.	0.	0.	0.	999.	999.	999.	999.
64	207006.	122.	0.	0.	0.	0.	0.	0.	0.	200.
65	207007.	300.	0.	100.	5.	2.	0.	0.	0.	0.
66	207008.	400.	20.	400.	5.	0.	0.	0.	0.	50.
67	208001.	250.	15.	250.	0.	0.	0.	0.	0.	20.
68	208002.	40.	0.	40.	3.	2.	0.	0.	0.	0.
69	208003.	150.	80.	80.	100.	4.	1.	0.	0.	0.
70	208004.	700.	10.	700.	0.	0.	0.	0.	0.	2.
71	208005.	250.	30.	100.	0.	2.	0.	0.	0.	0.
72	309001.	3200.	30.	30.	2.	0.	0.	0.	0.	0.
73	309002.	400.	300.	300.	150.	5.	0.	0.	0.	0.
74	309003.	500.	250.	250.	0.	0.	0.	0.	0.	0.
75	309004.	600.	30.	50.	3.	2.	0.	0.	0.	0.
76	309005.	350.	30.	250.	3.	0.	0.	0.	0.	0.
77	309006.	570.	200.	200.	20.	0.	0.	0.	0.	0.
78	309007.	400.	0.	200.	25.	0.	0.	0.	200.	0.
79	309008.	2000.	100.	0.	50.	2.	0.	0.	0.	1.
80	309009.	784.	100.	700.	40.	3.	0.	2.	300.	0.
							0.	0.	0.	0.

EXERCISE THREE
EXAMINING THE DATA 23-Feb-83 Page 3
File BIGTUNIS (Creation date = 02/04/83)
Subfile WEST SOFI CNTR NRTH UEND

Variable V215 # DRY FARM HOMES WITH GLASS WINDOWS

Mean	69.922	Std error	9.496	Std dev	83.326
Variance	6943.283	Kurtosis	1.724	Skewness	1.524
Range	350.000	Minimum	0.000	Maximum	350.000
Sum	5384.000				

Valid observations - 77 Missing observations - 3

EXERCISE THREE
EXAMINING THE DATA 23-Feb-83 Page 4

```
Transpace required:      130 words
      5 transformations
      0 RECODE values + LAG variables
     20 IF/COMPUTE operations

CPU time required:     2.28 seconds


                 16 FINISH

        Normal end of job
           16 control cards were processed
            0 errors were detected
```

```
/SS Patch System                                                                    23-Feb-83        Page   1
SPSS for DECSYSTEM-20, Version M, Release 9.1, May 1, 1982

                                Current Documentation for the SPSS Batch System
;Order from McGraw Hill:  SPSS, 2nd Ed. (principal text)        Order from SPSS Inc.:  SPSS Statistical Algorithms
                          SPSS Primer (brief intro to SPSS)                            KEYWORDS: The SPSS Inc. Newsletter
                          SPSS Update 7-9 (use w/SPSS,2nd for Rel. 7, 8, 9)
                          SPSS Introductory Guide: Basic Statistics and Operations
                          SPSS Primer (Brief Intro to SPSS)

Default SPACE allocation:     Allows for:          98 Transformations
WORKSPACE      17920 words                        394 RECODE values + LAG variables
TRANSPACE       2560 words                       1576 IF/COMPUTE operations

                1 RUN NAME       EXERCISE THREE
                2 GET FILE       BIGTUN.FIL

WARNING: Expected file BIGTUN   found file BIGTUNIS.


File BIGTUNIS has  497 variables

The subfiles are:

          No. of
Name      Cases

WEST         24
SOUT         11
CNTR         17
NRTH         19
JEND          9


CPU time required:     0.87 seconds


                3 COMPUTE          DWIND=(V215/V211)*100
                4 COMPUTE          DDOOR=(V216/V211)*100
                5 COMPUTE          DSTOVE=(V217/V211)*100
                6 COMPUTE          DSEW=(V218/V211)*100
                7 COMPUTE          DOVEN=(V223/V211)*100
                8 VAR LABELS       DWIND % DRYLAND FARMHOUSES WITH WINDOWS/DDOOR % DRYLAND
                9                  FARMS WITH DOORS/DSTOVE % DRYLAND FARMHOUSES WITH STOVES
               10                  DSEW % DRYLAND FARMHOUSES WITH SEWING MACHINES/DOVEN
               11                  % DRYLAND FARMHOUSES WITH OVENS
               12 ASSIGN MISSING   DWIND TO DOVEN(-1)
               13 TASK NAME        EXAMINING THE DATA
               14 LIST CASES       CASES=80/VARIABLES=ID V223 V224 DWIND TO DOVEN
               15 CONDESCRIPTIVE   V211

***** Given WORKSPACE allows for 1378 variables for CONDESCRIPTIVE problem *****
```

EXERCISE THREE
EXAMINING THE DATA 23-Feb-83 Page 2
File BIGTUNIS (Creation date = 02704783)
Subfile WEST SPEI CNTR NRTH JEND

Case	ID	V223	V224	DWIND	DDOOR	DSTOVE	DSFW	DOVEN
1	101001.	800.	0.	18.	50.	0.	0.	100.
2	101002.	6.	0.	0.	100.	0.	0.	100.
3	101003.	325.	0.	3.	8.	15.	2.	100.
4	101004.	300.	0.	43.	43.	1.	1.	43.
5	101005.	200.	0.	8.	40.	20.	0.	20.
6	101006.	450.	0.	13.	13.	3.	2.	100.
7	101007.	10.	0.	100.	100.	20.	0.	100.
8	101008.	1.	0.	23.	70.	0.	1.	0.
9	101009.	170.	0.	80.	100.	25.	2.	65.
10	102001.	40.	0.	0.	0.	0.	0.	65.
11	102002.	999.	999.	-1.	-1.	-1.	-1.	-1.
12	102003.	0.	0.	57.	57.	6.	1.	0.
13	102004.	600.	0.	33.	33.	7.	3.	100.
14	102005.	320.	0.	19.	78.	25.	1.	100.
15	102006.	319.	0.	47.	63.	9.	0.	100.
16	102007.	350.	0.	29.	20.	14.	1.	100.
17	102008.	250.	0.	4.	4.	0.	0.	100.
18	102009.	472.	0.	11.	32.	13.	0.	100.
19	102010.	0.	0.	5.	5.	0.	0.	0.
20	102011.	6.	0.	67.	67.	0.	2.	4.
21	102012.	484.	0.	10.	10.	14.	0.	100.
22	102013.	500.	0.	15.	15.	12.	1.	100.
23	102014.	0.	0.	25.	25.	0.	0.	0.
24	102015.	250.	0.	6.	6.	20.	2.	83.
25	103001.	0.	0.	7.	7.	3.	0.	0.
26	103002.	0.	0.	34.	34.	2.	1.	0.
27	103003.	200.	0.	8.	33.	1.	0.	56.
28	103004.	500.	0.	12.	12.	1.	0.	100.
29	103005.	0.	0.	43.	100.	1.	1.	0.
30	103006.	400.	0.	8.	8.	2.	0.	80.
31	103007.	0.	0.	44.	44.	1.	0.	0.
32	103008.	0.	0.	10.	10.	4.	0.	0.
33	103009.	15.	0.	-1.	-1.	-1.	-1.	-1.
34	103010.	100.	0.	25.	25.	0.	0.	23.
35	103011.	0.	0.	63.	75.	13.	1.	0.
36	104001.	80.	0.	8.	8.	3.	1.	20.
37	104002.	0.	0.	0.	0.	0.	0.	0.
38	104003.	100.	0.	0.	100.	4.	0.	20.
39	104004.	200.	0.	0.	100.	5.	0.	100.
40	104005.	500.	0.	1.	80.	3.	0.	100.
41	104006.	0.	0.	100.	100.	92.	5.	0.
42	104007.	350.	0.	1.	51.	1.	0.	59.
43	105001.	50.	0.	3.	29.	3.	0.	14.
44	105002.	200.	0.	30.	50.	15.	1.	100.
45	105003.	40.	0.	30.	60.	16.	4.	80.
46	105004.	33.	0.	100.	100.	21.	3.	100.

47	105005.	0.	0.	10.	10.	3.	1.	0
48	105006.	300.	0.	17.	67.	0.	0.	100.
49	105007.	300.	0.	0.	100.	0.	0.	100.
50	105008.	999.	999.	-1.	-1.	-1.	-1.	-1.
51	105009.	0.	0.	0.	100.	9.	1.	0.
52	105010.	300.	0.	13.	67.	0.	0.	100.
53	206001.	280.	0.	71.	71.	6.	6.	100.
54	206002.	0.	0.	0.	6.	0.	0.	0.
55	206003.	197.	0.	10.	10.	5.	1.	100.
56	206004.	475.	0.	0.	10.	0.	0.	100.
57	206005.	150.	0.	0.	100.	0.	0.	38.
58	206006.	0.	0.	22.	22.	0.	0.	0.
59	207001.	80.	0.	100.	100.	6.	0.	100.
60	207002.	900.	0.	1.	68.	1.	1.	62.
61	207003.	150.	999.	23.	23.	3.	0.	100.
62	207004.	999.	999.	-1.	-1.	-1.	-1.	-1.
63	207005.	200.	0.	0.	0.	0.	0.	50.
64	207006.	0.	0.	0.	0.	0.	0.	0.
65	207007.	300.	0.	0.	33.	2.	1.	100.
66	207008.	200.	0.	5.	100.	1.	0.	50.
67	208001.	250.	0.	6.	100.	0.	0.	100.
68	208002.	40.	0.	0.	100.	8.	5.	100.
69	208003.	15.	0.	53.	53.	67.	3.	10.
70	208004.	400.	0.	1.	100.	0.	0.	57.
71	208005.	0.	0.	12.	40.	0.	1.	0.
72	309001.	1000.	0.	1.	1.	0.	0.	31.
73	309002.	0.	0.	75.	75.	38.	1.	0.
74	309003.	500.	0.	50.	50.	0.	0.	100.
75	309004.	0.	0.	5.	8.	1.	0.	0.
76	309005.	350.	0.	9.	71.	1.	0.	100.
77	309006.	570.	0.	35.	35.	6.	0.	100.
78	309007.	400.	0.	0.	50.	6.	0.	100.
79	309008.	0.	0.	5.	10.	3.	0.	0.
80	309009.	200.	0.	13.	89.	5.	0.	25.

File SIGTUNIS (Creation date = 02/04/83)
Subfile WEST SBEI CNTR NRTH UEND

Variable V211 # OF DRY FARMERS HOUSES IN SUBDISTRICT

Mean	471.304	Std error	59.716	Std dev	530.769
Variance	281715.956	Kurtosis	14.912	Skewness	3.532
Range	3200.000	Minimum	0.000	Maximum	3200.000
Sum	37233.000				

Valid observations - 79 Missing observations - 1

```
Transpace required:        130 words
     5 transformations
     0 RECODE values + LAG variables
    20 IF/COMPUTE operations

CPU time required:     2.15 seconds


              16 FINISH

        Normal end of job
           16 control cards were processed
            0 errors were detected
```

Exercise 4: RELIABILITY AND THE COMPUTATION OF SCORES

Problem: How does one construct a score of a set of items that you believe are measuring the same thing?

Short answer: Test your hypothesis that a given set of items are measuring an identical concept by using the reliability test, discarding items that are unreliable, and then adding the remainder to make a score.

Reading: SPSS Update 7-9, page 248 ff. (However, much of this material is beyond the scope of this course.)

Output: As attached. The output consists of tests of reliability for three different sets of items, followed by a rerun after the information of the first output has been digested and used to delete items.

Commentary: If the social sciences had a number of standardized measures, such as temperature, weight, or even something as crude as the "wind-chill factor" that is often published in the newspapers, we would be very well off indeed. A few such standardized measures exist, but they usually do not apply to less developed areas and/or they refer to concepts that are not relevant. So it is necessary to construct your own measures, and this may require as much as 50 percent of the analysis time. In Exercise 3, we began the work of sorting items for a level-of-living scale, but computing that scale will require another exercise dealing with "z-scores," which are a means of standardizing the distribution. One uses z-scores in order to add items with different distributions, or, metaphorically, apples, oranges, pears, and bananas. More on that later.

Meanwhile, we introduce a simpler case where all the items have the same range of answers, either present or absent or one, two, three, and four. One does not usually need to standardize for these distributions. Furthermore, if you are sure that they all measure the same thing, you can simply add them up and divide

by the number of variables in order to have a "score." However, the assumption that all the items are measuring the same thing is a big one and must be examined carefully. One does that by looking at the items from a substantive point of view and by looking at the statistics generated by the RELIABILITY program.

Steps:

1. First select a set of items that you believe refer to the same dimension. In this case we have sets for "environmental problems," "women's work." and "restrictions on women." Some of these admit to a more theoretical interpretation, but in practical work nobody wants to hear about that, so we will ignore that problem. Note that each of these sets comes from a particular page of the questionnaire. In other words, we had in mind the possibility of using these sets for the calculation of indices. There is some redundancy in these items, which we might have weeded out had we had more time, but the sets in themselves do not represent redundancy. Just the opposite. They exemplify the principle in social measurement that one should never rely on a single item if one can avoid it.

2. Next look for faulty items. For example, V128, salinity of the water, would probably be present only in some parts of the region and therefore is a poor item. Salinity is associated with irrigation in dry areas, but the northern part of this region has little irrigation and much rainfall. So that item should be eliminated. V126 is probably ambiguous, although we can hold that judgment in reserve until we look at some of the statistics. Coming to the next set we note that most of the jobs that women hold are urban, with the exception of farm operator. V157 is probably ambiguous in that it comprehends too many different kinds of crafts, and V158 is even worse. That item might have been used had we coded the particular specifications, but evidently that was not done.

Finally, all items of the third set still look pretty good in the cool light of the office, but we can refine our judgement by looking at the statistics.

3. This exercise is intentionally asking you to compare three sets at a time. That is perhaps a little confusing, but it will give you a perspective on what you are likely to find. Accordingly, we scan the means, standard deviations, and number of cases for the three sets. Note that the number of cases has been reduced by the deletions of the towns, and there is one case with so much missing data that it could not be used. As for the means, you are looking for two things. You don't want any items with means below .10 or above .90 because they are not very useful for measurement unless you are really interested in extremes. One of these environmental problems items does hit .91, but we can retain it for now. The second thing you are looking for in these means is discrimination. You want items that cut off different proportions of the sample, and in this case the means run from .40 to .906, which is a good range.

The means for female employment are much lower, as one would expect in these male-dominated regions. V155, female driver, shows only .026 percent and may have to be dropped. On the other hand, V157, women employed as craft workers, is so high (76 percent) relative to the others as to make one suspicious.

The means for restrictions on women are all greater than one because the informants had four categories to choose from, and these were scored one, two, three, four. The fact that they are all about two is amazing, because some of these items were designed to reflect extremes in customs.

4. The next step is to scan the correlation matrices. What you would like to find is a set of uniformly high and positive (assuming that the question is worded in a positive way) correlations for all the items that you plan to keep. However, that is not the case on page 2. Thus, V123 and V128 introduce negative correlations, and it is pretty hard to justify these on the assumption that we are trying to measure ecological environmental problems. The correlations are rather low, anyway. On the other hand, it may be unreasonable to expect high correlations in this situation, because we are dealing with a semi-arid region where the problems are particularistic. For example, a given cluster of huts on the side of a ravine may have a problem with water erosion, but they may not suffer from remoteness from water. Clearly, this situation departs from the usual situation in survey analysis where you ask a single respondent a number of questions dealing with his "attitude" about this or that subject. So perhaps we should suspend judgment on this set of variables for the present.

Turning to page 5, we are reassured to see that the correlations are a little higher, but there are still some negative and zero correlations. The relationship of V155 and V154 yields a correlation of -.06, which is not desirable; but it is not clear which variable is causing the trouble. The interesting fact is that V155, which we have already proposed to drop on the grounds that it is so rare, actually correlates moderately with the first three variables. The other negative correlations are probably caused by V157 and V158, which we are going to drop anyway. So that matrix is looking pretty good, at least from the point of view of positive correlations.

On page 8 we have the kind of correlation matrix that one expects. There are no negative correlations, and the relationships are at least

moderately high.

5. Now we are ready to look at the reliability analysis, although it is not necessary to deal with every bit of output. Across the top of page 3 are the summary statistics, quite similar to those produced by CONDESCRIPTIVE. Now there is no need to guess what the mean correlation is. It has been calculated and comes to .145. The next number to look at is at the bottom of the table, the alpha coefficient. This is a general index of reliability, and the minimum is .50. So these alphas are on the low side, but perhaps it cannot be helped. The question is: Can we improve the reliability by dropping selected items? In order to decide that, look at the last column to see what the alpha would be if a particular item is deleted. It appears that deleting V123 or V126 would increase the alpha to .59, which is better than the other numbers in the column. On the other hand, that increase is no better than the overall alpha of .59, and it is not much of a gain, so one wonders whether to do it. But recall that V123 at least was probably responsible for some of those negative correlations.

The effect of the negative correlations shows up in the third column labeled "corrected item-total correlation." Note that the correlation between V123 and the whole cluster is .17, which is lower than the others. V126 is also low: .15. By the way, this relationship of the item to the total is the one that some professors use to decide whether a given test question should be used the following year. If a given question correlates highly with the class's total score, the assumption is that the individual item is a good one. But that is not so here. So those two have got to go. In addition, V128 must be dropped, as argued previously.

Page 6 shows the statistics for female jobs, and we immediately see that the alpha coefficient is even lower than before. The previous discussion suggested dropping all but the first four items, and the decision on V158 is confirmed by the increase in the alpha coefficient to .576 if that item is deleted. However, no such increase accrues from dropping any of the other items. Looking then at the middle column it appears that V151 and V158 have the lowest correlations. So we have an ambiguous situation, but probably the best rule to follow is to put the emphasis on the substantive aspects of the items. On that basis, we could include V151 through V155 on the grounds that they all reflect urban occupations, in contrast to the others.

Finally, one sees a really adequate scale on page 9. The standardized alpha is .85, and no single item is outstandingly bad, although we do note the low correlations of V159 and V167.

6. The second section of the output shows the effect of dropping some of the problematic items. On page B2 we have eliminated all of the negative correlations, but the standardized alpha listed on page 3 is almost identical with the previous one. All one can say is that this new version is more defensible, and that is not a negligible gain when it comes to writing up the material.

On page 5 we note immediately that there is still a negative correlation in the matrix due, probably, to the low proportion of women who drive cars or pickup trucks. Of course, all of those proportions are rather low, but that cannot be helped. The standardized alpha coefficient shows a slight improvement, and there are no other statistical "clinkers."

7. There is no need to rerun the scale of restrictions on women; it was good as it stood. As for the first two, it should be apparent that different

analysts might have made different judgments on which items to delete or, indeed, whether to use the scale at all. They are weaker than one would like, but then lower correlations are usual when dealing with rural areas, scales involving nonsocial items, and, of course, the measurement error that is inherent in a rush job such as this.

One final question that someone may wonder about is what, if anything, can be done about the missing case? If only a few items are missing, then a close approximation might be made on the basis of the other items. If, however, all of the data are missing for a given scale, about the only thing that can be done is to assign the overall average score or, alternatively, the most likely score from a neighboring subdistrict. For the moment, it is probably just as convenient and efficient to call it a missing case and delete it.

```
**********SYSLOG   ********,***R83.R01 ***FMT <00>, *****198 LINES *** 11:31.36PM  23 NOV 81  WARREN          SUSAN BENTLEY

B I STEP NAME                 SPSS     I
B I CPU SECONDS                 1.97 I
B I CPU ACCS (CPU+5.06)         5.68 I
B I TOTAL I/O TIME, SECS       28.21 I
B I REGION USED                152K I
B ***********************************************************************************************************
B * MSG: CLASS FOR JOB- A *I          USED     LIMIT          MIN:SEC                          REGION          TOTAL   I
B *                    *I          -------- --------        USED   LIMIT  $CHARGED   ----------------       EXECUTION I
B * ACCOUNT SAVE IF    *I LINES      261      3000        ------- ------ ---------   MAX USED     152K       CHARGE   I
B * ROUTING IN THIS CLASS *I CARDS      0       500   CPU    :02    2:00     $.19   LIMIT        160K   ----------- I
B * JOBID DO:      1.92  *I FILING      0       500   I/O    :26    5:00     $.55   CHARGE       $.75          $1.49 I
B ***********************************************************************************************************

0 JOB 6934-EF187     FROM R83.R01   12:15.49 MONDAY NOVEMBER 23, 1981
0                    //EF187    JOB Y6H$$$$,'SUSAN BENTLEY        ',PRTY=4                    DLVR=WRN
C                    /*LTR REGION=160K
C                    // EXEC SPSS
C                    //FT03F001  DD  UNIT=SYSDA,DISP=SHR,DSN=Y6H.BSOC
C                      INPUT STREAM DATASET OF 12 CARDS
C                    //

9 *-----------------------------------ACCOUNTING CONTROL RECORD (ACR) DISPLAY-----------------------------------*
B I Y6H   R S 213 HR              ACCOUNT NO.:  U964F7AA99    STATUS:   ACTIVE      ACCT BALANCE:     $168.89          I
C I DFLT RETPD: SYS        DSACCESS: SYS  DASD+CF PER DAY CHG:   $.13  DASD+CF TOT CHG:      $19.66                I
C I HASP TOTAL CHG:      $106.51                HASP CHG LIMIT:     $300.00                HASP BALANCE:     $193.49   I
B *-----------------------------------------------------------------------------------------------------------*
C //EF187    JOB Y6H$$$$,'SUSAN BENTLEY        ',PRTY=4                    DLVR=WRN
C // EXEC SPSS
C XXSPSS PROC SPACE=20K,TYPE1=800,TYPE2=2012,TYPE3=TRK,                    00000010
C XX      MAIN1=250,MAIN2=750,MAIN3=80,MAIN4=2250,                         00000020
C XX      SEC1=50,SEC2=190,LTB='SYS2.SPSS.LOAD',VERSION=F,                 00000030
C XX      ARCHIVE='DUMMY,'                                                 00000040
C ***                                                                  *** 00000050
C *** CURRENTLY ACCESSING THE INSTALLED VERSION OF SPSS 8.1 (05/20/80)*** 00000060
C *** PROBLEMS SHOULD BE REPORTED TO USER SERVICES AT 607-256-4981.   *** 00000070
C ***                                                                  *** 00000080
C *** PROCEDURE   DEFAULTS  ALLOW   SUFFICIENT   SCRATCH  SPACE   FOR *** 00000090
C *** PROCESSING UP TO  1500 CASES WITH 250 VARIABLES.  OR MORE CASES *** 00000100
C *** WITH  FEWER VARIABLES.   IF  MORE SPACE  IS  NEEDED, USERS  MAY *** 00000110
C *** INCREASE IT ON THE FT01, FT02, &  SORTWK JCL CARDS BY USING THE *** 00000120
C *** MAIN1, MAIN2, MAIN3, SEC1, SEC2, & TYPE3 SUBSTITUTION PARAMETERS*** 00000130
C ***                                                                  *** 00000140
C *** IF 'GET ARCHIVE', 'SAVE ARCHIVE', OR 'MERGE FILES' IS TO BE USED*** 00000150
C *** THEN CODE ARCHIVE=  ON THE EXEC CARD IN ORDER TO ACTIVATE THE   *** 00000160
C *** NECESSARY SCRATCH SPACE.                                         *** 00000170
C ***                                                                  *** 00000180
C *** FOR THE MAXI VERSION OF SPSS CODE THE FOLLOWING ON THE EXEC CARD*** 00000190
C ***                    // EXEC SPSS,VERSION=M                       *** 00000200
C XXSPSS     EXEC PGM=SPSS&VERSION,PARM=&SPACE                         00000210
C IEF653I SUBSTITUTION JCL - PGM=SPSSM,PARM=20K
C XXSTEPLTB  DD DSN=&LTB,DISP=SHR                                      00000220
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSS.LOAD,DISP=SHR
C XXFT01F001 DD UNIT=SYSSC,SPACE=(&TYPE1,(&MAIN1,&SEC1)),DSN=&&LABELS,  00000230
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(800,(250,50)),DSN=&LABELS,
C XX             DISP=(,DELETE)                                        00000240
C XXFT02F001 DD UNIT=SYSSC,SPACE=(&TYPE2,(&MAIN2,&SEC2)),DSN=&&DATASPSS, 00000250
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(2012,(750,190)),DSN=&DATASPSS,
C XX             DISP=(,DELETE)                                        00000260
C XXFT05F001 DD DUMMY=SYSIN                                            00000270
C XXFT06F001 DD SYSOUT=A                                               00000280
C XXFT10F001 DD DSN=SYS2.SPSS&VERSION..STAT,DISP=SHR                   00000290
C XXFT11F001 DD DSN=SYS2.SPSS&VERSION..STAT,DISP=SHR
```

A.

SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM

ORDER FROM MCGRAW-HILL: SPSS, 2ND ED. (PRINCIPAL TEXT)
SPSS PRIMER (BRIEF INTRO TO SPSS)
SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)

ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS
SPSS POCKET GUIDE, RELEASE 8
KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION.. ALLOWS FOR..
WORKSPACE 17920 BYTES 25 TRANSFORMATIONS
TRANSPACE 2560 BYTES 102 RECODE VALUES + LAG VARIABLES
412 IF/COMPUTE OPERATIONS

1 GET FILE FSOC

FILE FSOC W/S 409 VARIABLES

THE SUBFILES ARE..

NAME	CASES
[illegible]	[illegible]
[illegible]	11
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

CPU TIME REQUIRED.... [illegible] SECONDS

```
2 RUN NAME      RELIABILITY TESTS
3 SELECT IF     (NOTIID EQ 102001 OR 102002 OR 103011 OR 207004))
4 RELIABILITY   VARIABLES=V121 TO V129/
5               SCALE(ENVIRO)=V121 TO V129/
6 STATISTICS    1,3,8,9
```

****** METHOD 2 (COVARIANCE MATRIX) WILL BE USED FOR THIS ANALYSIS ******

****** 1276 BYTES OF SPACE REQUIRED FOR RELIABILITY ******

FILE RSOC (CREATION DATE = 11/04/81)
SUBFILE WEST SPET CNTR NPTH JEKD

* * * * * * * * * R E L I A B I L I T Y A N A L Y S I S F O R S C A L E (E N V I R O) * * * * * * * * *

1.	V121	DEEPENING EROSION OF RIVERS & GULLIES
2.	V122	ENLARGEMENT OF INUNDATED AREAS
3.	V123	ARRESTED GROWTH OF VEGETATION
4.	V124	REMOTENESS FROM PASTURES
5.	V125	DIFFICULTY IN FINDING FIREWOOD
6.	V126	WIND DAMAGE
7.	V127	REMOTENESS FROM WATER POINTS
8.	V128	SALINITY OF THE WATER
9.	V129	CONTAMINATION OF THE WATER

		MEANS	STD DEV	CASES
1.	V121	0.90667	0.29286	75.0
2.	V122	0.53333	0.50225	75.0
3.	V123	0.62667	0.48695	75.0
4.	V124	0.65333	0.47911	75.0
5.	V125	0.89333	0.31077	75.0
6.	V126	0.70667	0.45836	75.0
7.	V127	0.80000	0.40269	75.0
8.	V128	0.40000	0.49220	75.0
9.	V129	0.42667	0.49792	75.0

CORRELATION MATRIX

	V121	V122	V123	V124	V125	V126	V127	V128	V129
V121	1.00000								
V122	0.15925	1.00000							
V123	0.13140	0.16208	1.00000						
V124	0.15153	0.10483	0.24868	1.00000					
V125	0.18610	0.19625	-0.00011	0.29285	1.00000				
V126	0.09530	0.16045	0.10818	0.08451	0.06190	1.00000			
V127	0.18334	0.06682	0.16540	0.26616	0.36714	0.04393	1.00000		
V128	0.07405	0.21822	-0.04501	0.13725	0.01763	-0.07173	0.13608	1.00000	
V129	0.18411	0.26658	-0.00297	0.06193	0.21876	0.14132	0.29654	0.45126	1.00000

FILE RSOC (CREATION DATE = 11/04/81)
SUBFILE WEST SBET CNTR NRTH JEAD

* * * * * * * * * * R E L I A B I L I T Y A N A L Y S I S F O R S C A L E (E N V I R O) * * * * * * * * *

N OF CASES = 75.0

INTER-ITEM CORRELATIONS	MEAN	MINIMUM	MAXIMUM	RANGE	MAX/MIN	VARIANCE
	0.14513	-0.00611	0.45123	0.53934	-5.12127	0.01314

ITEM-TOTAL STATISTICS

	SCALE MEAN IF ITEM DELETED	SCALE VARIANCE IF ITEM DELETED	CORRECTED ITEM-TOTAL CORRELATION	SQUARED MULTIPLE CORRELATION	ALPHA IF ITEM DELETED
V121	5.04000	3.33622	0.28497	0.09168	0.56773
V122	5.41333	2.89441	0.33950	0.17113	0.54567
V123	5.32000	3.19351	0.17020	0.17745	0.59619
V124	5.29333	2.96685	0.32139	0.20800	0.55151
V125	5.05333	3.29441	0.29771	0.28158	0.56418
V126	5.24000	3.26595	0.15139	0.07701	0.59086
V127	5.14667	3.04577	0.36919	0.25046	0.54155
V128	5.54667	3.06198	0.24427	0.27747	0.57499
V129	5.52000	2.79351	0.41179	0.33090	0.52234

RELIABILITY COEFFICIENTS 9 ITEMS

ALPHA = 0.59231 STANDARDIZED ITEM ALPHA = 0.60443

```
TRANSPACE REQUIRED..      100 BYT
     1 TRANSFORMATIONS
     0 RECODE VALUES + LAG VARIABLES
    16 IF/COMPUTE OPERATIONS

CPU TIME REQUIRED..    0.52 SECONDS


                 7 RELIABILITY    VARIABLES=V151 TO V158/
                 8                SCALE(FEMALE)=V151 TO V158/
                 9 STATISTICS     1,3,8,9

 ****** METHOD 2 (COVARIANCE MATRIX) WILL BE USED FOR THIS ANALYSIS ******

 ******     1056 BYTES OF SPACE REQUIRED FOR RELIABILITY ******
```

Best Available Document

FILE RSOC (CREATION DATE = 11/04/81)
SUBFILE WEST SBEI CNTP METH JERD

* * * * * * * * * R E L I A B I L I T Y A N A L Y S I S F O R S C A L E (F E M A L E) * * * * * * * * *

1.	V151	WOMEN EMPLOYED AS REGISTERED MIDWIFE
2.	V152	WOMEN EMPLOYED AS REGISTERED NURSE
3.	V153	WOMEN SOCIAL SERVICE ASSISTANT
4.	V154	WOMEN SECRETARY,OFFICE WORKER, ETC
5.	V155	WOMEN DRIVER OF CAR OR PICKUP TRUCK
6.	V156	WOMEN EMPLOYED AS FARM OPERATOR
7.	V157	WOMEN EMPLOYED AS CRAFT WORKER
8.	V158	WOMEN EMPLOYED AS OTHER - SPECIFIED

		MEANS	STD DEV	CASES
1.	V151	0.00000	0.27312	75.0
2.	V152	0.05333	0.22621	75.0
3.	V153	0.17333	0.38108	75.0
4.	V154	0.13333	0.34222	75.0
	V155	0.02667	0.16219	75.0
	V156	0.21333	0.41242	75.0
	V157	0.76000	0.42996	75.0
	V158	0.36000	0.48323	75.0

CORRELATION MATRIX

	V151	V152	V153	V154	V155	V156	V157	V158
V151	1.00000							
V152	0.36746	1.00000						
V153	0.12464	0.36159	1.00000					
V154	0.17349	0.08146	0.33849	1.00000				
V155	0.25625	0.32903	0.36147	-0.06492	1.00000			
V156	0.08638	0.02124	0.27744	0.27447	0.11503	1.00000		
V157	-0.06444	0.13338	0.17485	0.12850	0.09301	0.21643	1.00000	
V158	-0.01638	-0.05439	0.02348	0.11440	-0.12414	0.15189	0.09626	1.00000

Best Available Document

FILE RSOC (CREATION DATE = 11/04/81)
SUBFILE WEST SPEI CNTR NFTH JEAD

* * * * * * * * * R E L I A B I L I T Y A N A L Y S I S F O R S C A L E (F E M A L E) * * * * * * * * *

OF CASES = 75.0

INTER-ITEM CORRELATIONS	MEAN	MINIMUM	MAXIMUM	RANGE	MAX/MIN	VARIANCE
	0.14204	-0.12414	0.36746	0.49160	-2.96003	0.01961

ITEM-TOTAL STATISTICS

	SCALE MEAN IF ITEM DELETED	SCALE VARIANCE IF ITEM DELETED	CORRECTED ITEM-TOTAL CORRELATION	SQUARED MULTIPLE CORRELATION	ALPHA IF ITEM DELETED
V151	1.72000	1.66378	0.17952	0.21679	0.51574
V152	1.74667	1.65117	0.27956	0.28227	0.49422
V153	1.62667	1.34523	0.42355	0.34633	0.42285
V154	1.66667	1.46847	0.33672	0.22951	0.46294
V155	1.77333	1.74523	0.21780	0.26252	0.51383
V156	1.58667	1.35387	0.35520	0.16862	0.44903
V157	1.04000	1.44432	0.22405	0.09999	0.50590
V158	1.44000	1.52000	0.09345	0.05322	0.57614

RELIABILITY COEFFICIENTS 8 ITEMS

ALPHA = 0.52825 STANDARDIZED ITEM ALPHA = 0.56980

CPU TIME REQUIRED.. 0.32 SECONDS

```
10 RELIABILITY   VARIABLES=V159 TO V165/
11               SCALE(RESTRICT)=V159 TO V165/
12 STATISTICS    1,3,8,9
```

****** METHOD 2 (COVARIANCE MATRIX) WILL BE USED FOR THIS ANALYSIS ******

****** 840 BYTES OF SPACE REQUIRED FOR RELIABILITY ******

FILL RSOC (CREATION DATE = 11/04/81)
SUBFILE WEST SUBT CNTR NPTH JEND

* * * * * * * * * RELIABILITY ANALYSIS FOR SCALE (RESTRICT) * * * * * * * * *

1.	V159	WOMEN STAY HOME MORE THAN MEN
2.	V160	WOMEN WEAR THE VEIL IN PUBLIC
3.	V161	WOMEN WEAR TRADITIONAL DRESS
4.	V162	WOMEN EAT APART FROM THE MEN
5.	V163	WOMEN DONT TALK WHEN VISITORS PRESENT
6.	V164	WOMEN SERVE MEAL & LEAVE
7.	V165	WOMEN SECLUDED BEFORE MARRY & BIRTH

		MEANS	STD DEV	CASES
1.	V159	2.46000	0.57797	75.0
2.	V160	2.25333	0.69929	75.0
3.	V161	2.17333	0.62327	75.0
4.	V162	2.22667	0.74568	75.0
5.	V163	2.18667	0.78316	75.0
6.	V164	2.26667	0.79812	75.0
7.	V165	2.20000	0.73521	75.0

CORRELATION MATRIX

	V159	V160	V161	V162	V163	V164	V165
V159	1.00000						
V160	0.56438	1.00000					
V161	0.47867	0.54900	1.00000				
V162	0.37124	0.71768	0.46677	1.00000			
V163	0.42632	0.55404	0.45883	0.64392	1.00000		
V164	0.24732	0.38773	0.47870	0.44424	0.61420	1.00000	
V165	0.08904	0.34695	0.39517	0.45847	0.45062	0.39347	1.00000

FILE RSOC (CREATION DATE = 11/04/81)
SUBFILE TEST SOFI CMTP NRTH JFRD

* * * * * * * * * * RELIABILITY ANALYSIS FOR SCALE (RESTRICT) * * * * * * * * * *

N OF CASES = 75.0

INTER-ITEM CORRELATIONS	MEAN	MINIMUM	MAXIMUM	RANGE	MAX/MIN	VARIANCE
	0.45913	0.08204	0.71768	0.62864	8.75990	0.01807

ITEM-TOTAL STATISTICS

	SCALE MEAN IF ITEM DELETED	SCALE VARIANCE IF ITEM DELETED	CORRECTED ITEM-TOTAL CORRELATION	SQUARED MULTIPLE CORRELATION	ALPHA IF ITEM DELETED
V159	13.50667	11.00036	0.47833	0.42423	0.85149
V160	13.53333	9.65766	0.70806	0.63595	0.81940
V161	13.61333	10.29441	0.63539	0.45264	0.83122
V162	13.56000	9.38486	0.71863	0.62468	0.81691
V163	13.60000	9.13514	0.73418	0.59011	0.81396
V164	13.52000	9.71243	0.58207	0.43998	0.83905
V165	13.50667	10.40793	0.47472	0.32654	0.85278

RELIABILITY COEFFICIENTS 7 ITEMS

ALPHA = 0.85329 STANDARDIZED ITEM ALPHA = 0.85345

Best Available Document

```
*****************SYSMSG    ************RM3.RD1 ***FMT <00>* *****198 LINES ***  7:36.04AM   2 DEC 81  WARREN                SUSAN BENTLEY
B | STEP NAME                    SPSS  |
B | CPU SECONDS                    1.40 |
B | CPU MIN (CPU*5.00)             7.00 |
B | TOTAL I/O TIME, SECS          24.74 |
B | REGION USED                    152K |
B ***************************************************************************************************************
B * BEST CLASS FOR JOB- A *|          USED     LIMIT          MIN:SEC                                  REGION          TOTAL  |
B *                       *|          --------- --------      USED   LIMIT  $CHARGED                ----------------   EXECUTION |
B * AMOUNT SAVED BY       *| LINES       167     3000       ------- ------- ---------                MAX USED     152K    CHARGE |
B * RUNNING IN THIS CLASS *| CARDS         0      500   CPU    :01    2:00     $.13                 LIMIT        160K  ---------- |
B * WOULD BE:        $.02 *| FILING        0      500   I/O    :25    5:00     $.48                 CHARGE       $.63       $1.24 |
B ***************************************************************************************************************

B ---GB 7650-EM996     FROM RM3.RD1   12:37.50 TUESDAY DECEMBER 1, 1981
B                      //EM996    JOB Y6H$$$$,'SUSAN BENTLEY        ',PRTY=4                          DLVR=WRN
C                      /*LIM REGION=160K
C                      // EXEC S SS
C                      //FT03F001  DD  UNIT=SYSDA,DISP=SHR,DSN=Y8H.RSOC
C                        INPUT STREAM DATASET OF 9 CARDS
C                      //

B *-----------------------------------ACCOUNTING CONTROL RECORD (ACR) DISPLAY-----------------------------------*
B | Y8H   R S 213 HR                ACCOUNT NO.:  U964F7AA99    STATUS:   ACTIVE      ACCT BALANCE:     $157.13      |
C | DFLT NETFE: SYS          DSACCESS: SYS  DASD+CF PER DAY CHG:   $.13  DASD+CF TOT CHG:      $22.14            |
C | HASP TOTAL CHG:      $114.19                HASP CHG LIMIT:     $300.00                HASP BALANCE:     $185.81 |
B *-----------------------------------------------------------------------------------------------------------*
C //EM996    JOB Y6H$$$$,'SUSAN BENTLEY        ',PRTY=4                                  DLVR=WRN
C // EXEC SPSS
C XXSPSS PROC SPACE=20K,TYPE1=600,TYPE2=2012,TYPE3=TRK,                                  00000010
C XX       MAIN1=250,MAIN2=750,MAIN3=80,MAIN4=2250,                                      00000020
C XX       SEC1=50,SEC2=190,LIB='SYS2.SPSS.LOAD',VERSION=H,                              00000030
C XX       ARCHIVE='DUMMY,'                                                              00000040
C ***                                                                                    00000050
C *** CURRENTLY ACCESSING THE INSTALLED VERSION OF SPSS 8.1 (05/20/80)*** 00000060
C *** PROBLEMS SHOULD BE REPORTED TO USER SERVICES AT 607-256-4981.   *** 00000070
C ***                                                                 *** 00000080
C *** PROCEDURE   DEFAULTS  ALLOW   SUFFICIENT   SCRATCH   SPACE   FOR *** 00000090
C *** PROCESSING UP TO  1500 CASES WITH 250 VARIABLES,  OR MORE CASES *** 00000100
C *** WITH  FEWER VARIABLES.   IF  MORE SPACE  IS  NEEDED, USERS  MAY *** 00000110
C *** INCREASE IT ON THE FT01, FT02, &  SORTWK JCL CARDS BY USING THE *** 00000120
C *** MAIN1, MAIN2, MAIN3, SEC1, SEC2, & TYPE3 SUBSTITUTION PARAMETERS*** 00000130
C ***                                                                 *** 00000140
C *** IF 'GET ARCHIVE', 'SAVE ARCHIVE', OR 'MERGE FILES' IS TO BE USED*** 00000150
C *** THEN CODE ARCHIVE=  ON THE EXEC CARD IN ORDER TO ACTIVATE THE   *** 00000160
C *** NECESSARY SCRATCH SPACE.                                        *** 00000170
C ***                                                                 *** 00000180
C *** FOR THE MAXI VERSION OF SPSS CODE THE FOLLOWING ON THE EXEC CARD*** 00000190
C ***                    // EXEC SPSS,VERSION=M                      *** 00000200
C XXSPSS     EXEC PGM=SPSS&VERSION,PARM=&SPACE                                           00000210
C IEF653I SUBSTITUTION JCL - PGM=SPSSH,PARM=20K
C XXSTEPLIB  DD DSN=&LIB,DISP=SHR                                                        00000220
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSS.LOAD,DISP=SHR
C XXFT01F001 DD UNIT=SYSSC,SPACE=(&TYPE1,(&MAIN1,&SEC1)),DSN=&&LABELS,                   00000230
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(600,(250,50)),DSN=&LABELS,
C XX            DISP=(,DELETE)                                                           00000240
C XXFT02F001 DD UNIT=SYSSC,SPACE=(&TYPE2,(&MAIN2,&SEC2)),DSN=&&DATASPSS,                 00000250
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(2012,(750,190)),DSN=&DATASPSS,
C XX            DISP=(,DELETE)                                                           00000260
C XXFT05F001 DD DUMMY=SYSIN                                                              00000270
C XXFT06F001 DD SYSOUT=A                                                                 00000280
C XXFT10F001 DD DSN=SYS2.SPSS&VERSION..STAT,DISP=SHR                                     00000290
```

AB

New reliability on V121+ and on V151+

SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM

ORDER FROM MCGRAW-HILL: SPSS, 2ND ED. (PRINCIPAL TEXT)
SPSS PRIMER (BRIEF INTRO TO SPSS)
SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)

ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS
SPSS POCKET GUIDE, RELEASE 8
KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION..
WORKSPACE 17920 BYTES
TRANSPACE 2560 BYTES

ALLOWS FOR..
25 TRANSFORMATIONS
102 RECODE VALUES + LAG VARIABLES
412 IF/COMPUTE OPERATIONS

```
  1 GET FILE      RSCC
```

FILE RSCC HAS 409 VARIABLES

THE SUBFILES ARE..

NAME	NO OF CASES
WEST	24
SBEI	11
CNTR	17
NRTH	19
JEND	9

CPU TIME REQUIRED.. 0.45 SECONDS

```
  2 RUN NAME        RELIABILITY TESTS
  3 SELECT IF       (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
  4 RELIABILITY     VARIABLES=V121,V122,V124,V125,V127,V129/
  5                 SCALE(ENVIRO)=V121,V122,V124,V125,V127,V129/
  6 STATISTICS      1,3,8,9
```

****** METHOD 2 (COVARIANCE MATRIX) WILL BE USED FOR THIS ANALYSIS ******

****** 640 BYTES OF SPACE REQUIRED FOR RELIABILITY ******

FILE RSOC (CREATION DATE = 11/04/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * * * * R E L I A B I L I T Y A N A L Y S I S F O R S C A L E (E N V I R O) * * * * * * * * *

1.	V121	DEEPENING EROSION OF RIVERS & GULLIES
2.	V122	ENLARGEMENT OF INUNDATED AREAS
3.	V124	REMOTENESS FROM PASTURES
4.	V125	DIFFICULTY IN FINDING FIREWOOD
5.	V127	REMOTENESS FROM WATER POINTS
6.	V129	CONTAMINATION OF THE WATER

		MEANS	STD DEV	CASES
1.	V121	0.90667	0.29286	75.0
2.	V122	0.53333	0.50225	75.0
3.	V124	0.65333	0.47911	75.0
4.	V125	0.89333	0.31077	75.0
5.	V127	0.80000	0.40269	75.0
6.	V129	0.42667	0.49792	75.0

CORRELATION MATRIX

	V121	V122	V124	V125	V127	V129
V121	1.00000					
V122	0.15925	1.00000				
V124	0.15153	0.10483	1.00000			
V125	0.18610	0.19625	0.29285	1.00000		
V127	0.18334	0.06682	0.26616	0.36714	1.00000	
V129	0.18411	0.26658	0.06193	0.21076	0.29654	1.00000

RELIABILITY TESTS

12/02/81 PAGE 3

FILE RSOC (CREATION DATE = 11/04/81)
SUBFILE WEST SDEI CNTR NRTH JEND

* * * * * * * * * R E L I A B I L I T Y A N A L Y S I S F O R S C A L E (E N V I R O) * * * * * * * * *

N OF CASES = 75.0

INTER-ITEM CORRELATIONS	MEAN	MINIMUM	MAXIMUM	RANGE	MAX/MIN	VARIANCE
	0.19961	0.06193	0.36714	0.30521	5.92811	0.00727

ITEM-TOTAL STATISTICS

	SCALE MEAN IF ITEM DELETED	SCALE VARIANCE IF ITEM DELETED	CORRECTED ITEM-TOTAL CORRELATION	SQUARED MULTIPLE CORRELATION	ALPHA IF ITEM DELETED
V121	3.30667	1.75604	0.28368	0.08103	0.54638
V122	3.68000	1.49081	0.26803	0.10795	0.56079
V124	3.56000	1.52000	0.26446	0.12647	0.55536
V125	3.32000	1.62595	0.42832	0.20991	0.49839
V127	3.41333	1.51604	0.38701	0.22269	0.49798
V129	3.78667	1.41333	0.33848	0.16517	0.51919

RELIABILITY COEFFICIENTS 6 ITEMS

ALPHA = 0.57483 STANDARDIZED ITEM ALPHA = 0.59942

```
TRANSPACE REQUIRED..      100 BYTES
     1 TRANSFORMATIONS
     0 RECODE VALUES + LAG VARIABLES
    16 IF/COMPUTE OPERATIONS

CPU TIME REQUIRED..    0.42 SECONDS


                 7 RELIABILITY    VARIABLES=V151 TO V155/
                 8                SCALE(FEMALE)=V151 TO V155/
                 9 STATISTICS     1,3,8,9

****** METHOD 2 (COVARIANCE MATRIX) WILL BE USED FOR THIS ANALYSIS ******

******      480 BYTES OF SPACE REQUIRED FOR RELIABILITY ******
```

* * * * * * * * * R E L I A B I L I T Y A N A L Y S I S F O R S C A L E (F E M A L E) * * * * * * * * *

1.	V151	WOMEN EMPLOYED AS REGISTERED MIDWIFE
2.	V152	WOMEN EMPLOYED AS REGISTERED NURSE
3.	V153	WOMEN SOCIAL SERVICE ASSISTANT
4.	V154	WOMEN SECRETARY,OFFICE WORKER, ETC
5.	V155	WOMEN DRIVER OF CAR OR PICKUP TRUCK

		MEANS	STD DEV	CASES
1.	V151	0.00000	0.27312	75.0
2.	V152	0.05333	0.22621	75.0
3.	V153	0.17333	0.38108	75.0
4.	V154	0.13333	0.34222	75.0
5.	V155	0.02667	0.16219	75.0

CORRELATION MATRIX

	V151	V152	V153	V154	V155
V151	1.00000				
V152	0.36746	1.00000			
V153	0.12464	0.36159	1.00000		
V154	0.17349	0.08146	0.33849	1.00000	
V155	0.25625	0.32903	0.36147	-0.06492	1.00000

N OF CASES = 75.0

INTER-ITEM CORRELATIONS	MEAN	MINIMUM	MAXIMUM	RANGE	MAX/MIN	VARIANCE
	0.23290	-0.06492	0.36746	0.43238	-5.65997	0.02095

ITEM-TOTAL STATISTICS

	SCALE MEAN IF ITEM DELETED	SCALE VARIANCE IF ITEM DELETED	CORRECTED ITEM-TOTAL CORRELATION	SQUARED MULTIPLE CORRELATION	ALPHA IF ITEM DELETED
V151	0.38667	0.56468	0.30815	0.19403	0.53095
V152	0.41333	0.57009	0.42303	0.25967	0.48378
V153	0.29333	0.39928	0.45943	0.32190	0.43442
V154	0.33333	0.52252	0.25492	0.19180	0.57471
V155	0.44000	0.65514	0.32116	0.24498	0.54345

102

FILE RSOC (CREATION DATE = 11/04/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * * * * R E L I A B I L I T Y A N A L Y S I S F O R S C A L E (F E M A L E) * * * * * * * * *

RELIABILITY COEFFICIENTS 5 ITEMS

ALPHA = 0.57353 STANDARDIZED ITEM ALPHA = 0.60286

CPU TIME REQUIRED.. 0.27 SECONDS

```
10 FINISH          (SPSS GENERATED)

NORMAL END OF JOB.
   10 CONTROL CARDS WERE PROCESSED.
    0 ERRORS WERE DETECTED.
```

EXERCISE 5 GUTTMAN SCALES

Problem: How can you combine dichotomous items into a scale?

Brief Answer: You use the Guttman or cumulative scaling technique.

Reading: SPSS Manual, Ch. 26; Baseline, p. 99-103.

Output: Sets A, B, and C.

Commentary: On the one hand, the informant survey methodology calls for many simple yes or no questions, but on the other it demands continuous variables so that you can do correlational or regression analysis. A standard technique for moving from the dichotomous qualitative data that the informants give you to a continuous variable is the technique of Guttman scaling. The basic assumption of this technique of scale construction is that certain attributes typically show up in a cumulative pattern. For example, you would expect someone who can do a calculus problem to be able to do a problem in trigonometry, in algebra, long division, and addition. On the other hand, the person who has only studied algebra is unlikely to be able to do the more "difficult" problems. There are exceptions to this general pattern, of course, but it is so likely that if you took the scores of 100 students on such a heterogenous test and shuffled them, you could probably find such a cumulative pattern. Likewise, if you find a village that has a market place, it is likely that it has a drug store, a hardware store, and telephone and postal service. That, at least, has been the assumption in much community research, and it seems to work on the whole.

In Tunisia we collected a wide range of dichotomous attributes such as this, but the initial attempts to construct a scale were unsuccessful. After examining the pool of items, it finally dawned on us that there were two basic dimensions involved: the local process of specialization illustrated by retail stores and second, the regional or government-supported institutions such as the postal and telephone service. Once this distinction was clear, we were able to find a workable scale for the regional items, and we ended up using a

z-score technique for combining a number of continuous variables reflecting local specialization.

In the course of this analysis, the question of how to handle the variable "number of retail stores" came up. We could have simply added it to the composite z-score for local institutions, but it seemed too important for this, so we kept it separate. It constitutes something of a validator, or at any rate, it is a useful variable to use by itself. It would also have been possible to dichotomize it by dividing it at the mean or mode, but that is simply "losing information" because you already have a continuous variable. So remember that Guttman scaling is appropriate for dichotomies or possibly trichotomies. If you have continuous variables, there is no gain in chopping them up so that you can make a scale. That misses the whole point of using the Guttman technique.

Steps:

1. Select a pool of items that reflect a single dimension of social structure, in this case regional or at least government-supported institutions. Eliminate ambiguous items, items that aren't institutions and, most importantly, items that don't fit the concept. Select a range of items that will cut off differing segments of the sample. If all the items occur in, say, 60 percent of the sample, they will not be useful as a measuring rod. Find items that appear in 10 percent of the sample, 20 percent, etc.

What you are attempting to find is represented to some extent in the printout in the C section of the output. Imagine that this listing of items is arranged so that those subdistricts that have all seven items are at the top, the subdistricts with six items are grouped next, and so on down the line. As you see, there will be some "error types," but that is always the case, and there is a way to handle this. What you looking for, generally speaking, is a "stair-step pattern." Note also that once the data are organized

this way, it is possible to compress them down as shown in the table in Baseline, page 128.

2. However, when you start the scaling process, you are not so sure how everything will fall out so you might get something like the output on page A2. Note first of all that each of the items is dichotomized. The program does this automatically if you specify a division point of "1.00." In other words, all these items are converted to the "presence of." V319 is put in there on purpose as an example of an item that shouldn't have been there in the first place. Roman ruins can hardly be defended as a contemporary institution, whether local or regional.

As you see from the numbers at the base of the "scalogram" output, the program arranges the items according to frequency. Thus, there are only eight cases of V130, presence of an urban water installation. That consitutes 11 percent of the sample, but as the last row shows, that particular item accounts for five errors. In other words, there are more communities "out of place" on this item than there are in the right place. Notice that there are two kinds of errors, and these are labeled in the scalogram. You can have an error when the item is present in the "0" column, and you can also have an error when it is absent in the "1" column. By the way, this particular output format is considered quite puzzling and should be revised, but SPSS has not gotten around to doing it.

At this point you might well ask isn't there some overall coefficient that will tell you whether the scale is minimally acceptable? Indeed there is, as shown on page 3. The coefficient that is used around here is the "coefficient of scalability" which in this case is .4967. The acceptable minimum is .60 so this scale is flawed in some way. The problem is, which item or items must be dropped to improve it? Alternatively, you might be able

to find other items that work better. We already know that V130 has to go, because it does more harm than good, and the other obvious candidate is V319, Roman ruins. That item introduces a total of 11 errors, and 10 of these are in the "0" column which has only 13 cases in all.

There is another problem with V002, presence of a primary school, because it introduces 3 errors, but doesn't discriminate any segment of the sample. Actually, it was included intentionally on the assumption that it would be a "threshold" or defining item. That is, it was assumed that all subdistricts would have at least one primary school; and while an item like that does not contribute to your measurement efficiency, it does provide a "floor item." The only problem with it is that it introduces 3 errors. However, it turned out that in all 3 cases a school was being built, so in the next run we recoded these 3 as "present." Thus, 100 percent of the cases would have schools. We also made another correction with respect to V328, presence of permanent markets. On the basis of our own observations and discussions with the interviewers, it became clear that some of the respondents misunderstood the question about permanent markets. What we wanted to know was whether there was a market structure, but some of them interpreted the question to mean whether the market met regularly. So we recoded those cases, too. However, that was all the additional information we had, so we simply had to live with the rest of the data, even though it is virtually certain that similar misunderstandings are imbedded in these facts.

Page B2 shows the effect of cleaning up the problems in the first scalogram. V328 is now the highest item, although it still is weak in that it accounts for only 5 percent of the sample. All the other items "work" in the sense that there are fewer errors than there are positive responses, but the fact is that there are still a quite a lot of errors in this scale. The

coefficient of scalability has increased to .58, but that is still less than the minimum. Page B2 actually reflects a number of trials, but they are not reproduced here. This is the best scale that emerged from this set of items, which is puzzling, because the same set as listed in Baseline gave a coefficient of scalability of .62. The difference is that we used an older SPSS program, which counted errors in a slightly different way.

Page B3 shows all the coefficients that are available in this program, although we are only using one of them. B3 also shows a correlation matrix based on Yule's Q, which is the measure of correlation for dichotomous items like these. Also shown is the biserial correlation, which is the relationship of a particular item to the whole scale. These correlation coefficients are useful in evaluating the items, but they do not take the place of proper conceptualization and inspection of the error patterns.

3. The information needed for studying the error patterns is shown in the listing in C3. Note that the control cards in this case require the specific recoding of the variables so that the 0-1 pattern is listed. Such specific recoding was not necessary, or was done automatically, in the Guttman scaling program.

As is evident from my marks, there are a lot of error types in this scalogram. The question is whether something can be done about some of them. The most serious are marked with a longer dash. Subdistrict 105002 has electricity, but then it lacks the next three "lower" items. What probably happened is that this subdistrict was close to the electric line or was located next to a large subdistrict with electricity. So it received its power connection somewhat accidentally. Field investigation might reveal similar explanations for other "errors." This word is in quotation marks because, of course, they are only errors from a statistical point of view. The electric lines are real enough.

4. When you finally have the best scalogram possible, the actual scale consists of a count of the "1's". But this is a different count from the scores we made for ecological problems. This score has an underlying structure in the sense that the technique insures "undimensionability," i.e., a single dimension. In contrast, a simple count of items, even if you believe they fit some concept, could muddle up several dimensions.

SPSS BATCH SYSTEM

SPSS FOR IBM/360, VERSION H, RELEASE 8.1, MAY 20, 1980

CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM

ORDER FROM MCGRAW-HILL:
SPSS, 2ND ED. (PRINCIPAL TEXT)
SPSS PRIMER (BRIEF INTRO TO SPSS)
SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)

ORDER FROM SPSS INC.:
SPSS STATISTICAL ALGORITHMS
SPSS POCKET GUIDE, RELEASE 8
KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION.. ALLOWS FOR..
WORKSPACE 17900 BYTES 25 TRANSFORMATIONS
TRANSPACE 25,0 BYTES 102 RECODE VALUES + LAG VARIABLES
412 IF/COMPUTE OPERATIONS

```
1 GET FILE      RSOC
```

FILE RSOC HAS 409 VARIABLES

THE SUBFILES ARE..

NAME	NO OF CASES
WEST	24
SOEI	11
CNTR	17
NRTH	19
JEWO	5

CPU TIME REQUIRED.. 0.64 SECONDS

```
2 SELECT IF      (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
3 *RECODE        V002(1 THRU HI=1)
4 *RECODE        V130(1 THRU HI=1)
5 GUTTMAN SCALE  GUTSCALE=V002(1) V130(1) V249(1) V302(1) V303(1) V319(1) V321(1)
6                V328(1)
7 STATISTICS     ALL
```

GUTMAN SCALE PROBLEM REQUIRES 680 BYTES WORKSPACE *****

FILE RSEC (CREATION DATE = 12/09/81)
SUBFILE WEST SHEI CATR NRTH JEND

* * * * * * * * * * GUTTMAN SCALE (GUTSCALE) USING * * * * * * * * * *

V002	# OF PRIMARY SCHOOLS	DIVISION POINT =	1.00
V150	# OF URBAN WATER AGENCIY INSTALLATIONS	DIVISION POINT =	1.00
V249	PRESENCE OF ELECTRICIYT IN SUBDISTRICT	DIVISION POINT =	1.00
V302	PRESENCE OF TELEPHONES	DIVISION POINT =	1.00
V303	PRESENCE OF POSTAL BOXES	DIVISION POINT =	1.00
V319	PRESENCE OF ROMAN RUINS	DIVISION POINT =	1.00
V321	PRESENCE OF DISPENSARIES	DIVISION POINT =	1.00
V329	PRESENCE OF PERMANENT MARKETS	DIVISION POINT =	1.00

* * * * * * * * * * * RESP = 1 FOR VALUES EQUAL TO DIVISION POINT AND ABOVE * * * * * * * * * * *

ITEM..	V150		V329		V249		V321		V303		V302		V319		V002		
RESP..	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	TOTAL
GUTTMAN SCALE																	
8	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	3
	ERR																
7	1	2	1	2	0	3	0	3	0	3	0	3	1	2	0	3	3
		ERR															
6	2	2	7	0	1	6	0	7	1	6	0	7	0	7	0	7	7
				ERR													
5	12	0	12	0	6	6	4	8	0	12	1	11	1	11	0	12	12
							ERR										
4	14	0	12	2	11	3	7	7	7	7	2	12	3	11	0	14	14
									ERR								
3	17	1	17	1	16	2	15	3	13	5	8	10	4	14	0	18	18
											ERR						
2	15	0	15	0	14	1	15	0	15	0	13	2	1	14	2	13	15
													ERR				
1	4	0	4	0	4	0	4	0	4	0	4	0	3	1	1	3	4
															ERR		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SUMS	65	8	68	8	52	24	45	31	40	36	28	48	13	63	3	73	76
PCTS	89	11	89	11	68	32	59	41	53	47	37	63	17	83	4	96	
ERRORS	0	5	1	3	1	12	4	10	8	5	11	2	10	1	3	0	76

76 CASES WERE PROCESSED
0 (0% 0.0 PCT) WERE MISSING

STATISTICS..

COEFFICIENT OF REPRODUCIBILITY = 0.8750
MINIMUM MARGINAL REPRODUCIBILITY = 0.7516
PERCENT IMPROVEMENT = 0.1234
COEFFICIENT OF SCALABILITY = 0.4967

YULE'S Q..

	V002	V130	V249	V302	V303	V319	V321	V328
V002	1.0000	1.0000	-0.0417	0.5667	1.0000	-1.0000	1.0000	1.0000
V130	1.0000	1.0000	0.7857	0.6435	0.5833	0.2000	0.8554	0.8824
V249	-0.0417	0.7857	1.0000	0.7327	0.7000	-0.1828	0.5709	0.6225
V302	0.5667	0.6435	0.7327	1.0000	0.7870	-0.1613	0.5988	0.3000
V303	1.0000	0.5833	0.7000	0.7870	1.0000	0.0294	0.6881	0.3309
V319	-1.0000	0.2000	-0.1828	-0.1613	0.0294	1.0000	0.0585	-0.2667
V321	1.0000	0.8554	0.5709	0.5988	0.6881	0.0585	1.0000	0.8554
V328	1.0000	0.8824	0.6225	0.3000	0.3309	-0.2667	0.8554	1.0000
BISERIAL CORR SCALE-ITEM	0.3276	0.6919	0.5658	0.5315	0.6289	-0.0605	0.6362	0.4995

SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM

ORDER FROM MCGRAW-HILL: SPSS, 2ND ED. (PRINCIPAL TEXT)
SPSS PRIMER (BRIEF INTRO TO SPSS)
SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)

ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS
SPSS POCKET GUIDE, RELEASE 8
KEYWORDS: THE SPSS INC. NEWSLETTER

```
DEFAULT SPACE ALLOCATION..     ALLOWS FOR..     25 TRANSFORMATIONS
WORKSPACE      17920 BYTES                     102 RECODE VALUES + LAG VARIABLES
TRANSPACE       2560 BYTES                     412 IF/COMPUTE OPERATIONS

          1 GET FILE       RSOC


                           FILE RSOC     HAS 409 VARIABLES

                           THE SUBFILES ARE..

                                    NO OF
                           NAME     CASES

                           WEST       24
                           SBEI       11
                           CNTR       17
                           NRTH       15
                           JEND        9


CPU TIME REQUIRED..     0.50 SECONDS


          2 SELECT IF      (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
          3 COMPUTE        PERMARKT=0
          4 IF             (ID EQ 102002)PERMARKT=1
          5 IF             (ID EQ 103011)PERMARKT=1
          6 IF             (ID EQ 104006)PERMARKT=1
          7 IF             (ID EQ 206004)PERMARKT=1
          8 IF             (ID EQ 207004)PERMARKT=1
          9 IF             (ID EQ 208003)PERMARKT=1
         10 IF             (ID EQ 309008)PERMARKT=1
         11 *RECODE        V002(0 THRU HIGHEST=1)
         12 *RECODE        V130(1 THRU HIGHEST=1)
         13 GUTTMAN SCALE  GUTSCALE=V002(1) V099(1) V249(1) V302(1) V303(1) V321(1)
         14                PERMARKT(1)
         15 STATISTICS     ALL

GUTMAN SCALE PROBLEM REQUIRES     532 BYTES WORKSPACE *****
```

114

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SHEI CNTR NRTH JEND

* * * * * * * * * G U T T M A N S C A L E (G U T S C A L E) U S I N G * * * * * * * * * *

V002	# OF PRIMARY SCHOOLS	DIVISION POINT =	1.00
V099	PRESENCE OF A MARKET IN SUBDISTRICT	DIVISION POINT =	1.00
V249	PRESENCE OF ELECTRICITYT IN SUBDISTRICT	DIVISION POINT =	1.00
V302	PRESENCE OF TELEPHONES	DIVISION POINT =	1.00
V303	PRESENCE OF POSTAL BOXES	DIVISION POINT =	1.00
V321	PRESENCE OF DISPENSARIES	DIVISION POINT =	1.00
PERMARKT		DIVISION POINT =	1.00

* * * * * * * * * * * RESP = 1 FOR VALUES EQUAL TO DIVISION POINT AND ABOVE * * * * * * * * * * *

ITEM..	PERMARKT		V099		V249		V321		V303		V302		V002		TOTAL
RESP..	0	1	0	1	0	1	0	1	0	1	0	1	0	1	
GUTSCALE 7	0	4	0	4	0	4	0	4	0	4	0	4	0	4	4
6	3	0	0	3	0	3	0	3	0	3	0	3	0	3	3
5	7	0	5	2	1	6	0	7	1	6	0	7	0	7	7
4	14	0	14	0	7	7	5	9	1	13	1	13	0	14	14
3	16	0	12	4	14	2	10	6	9	7	3	13	0	16	16
2	16	0	15	1	14	2	14	2	13	3	8	8	0	16	16
1	16	0	16	0	16	0	16	0	16	0	16	0	0	16	16
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SUMS	72	4	62	14	52	24	45	31	40	36	28	48	0	76	76
PCTS	95	5	82	18	68	32	59	41	53	47	37	63	0	100	
ERRORS	0	0	0	7	1	11	5	8	11	3	12	0	0	0	58

76 CASES WERE PROCESSED
0 (OR 0.0 PCT) WERE MISSING

STATISTICS..

COEFFICIENT OF REPRODUCIBILITY = 0.8910
MINIMUM MARGINAL REPRODUCIBILITY = 0.7425
PERCENT IMPROVEMENT = 0.1485
COEFFICIENT OF SCALABILITY = 0.5766

YULE'S Q..

	V002	V099	V249	V302	V303	V321	PERMARKT
V002	1.0000	99.0000	99.0000	99.0000	99.0000	99.0000	99.0000
V099	99.0000	1.0000	0.5862	0.4249	0.2364	0.7701	1.0000
V249	99.0000	0.5862	1.0000	0.7327	0.7000	0.5789	1.0000
V302	99.0000	0.4249	0.7327	1.0000	0.7870	0.5988	1.0000
V303	99.0000	0.2364	0.7000	0.7870	1.0000	0.6881	1.0000
V321	99.0000	0.7701	0.5789	0.5988	0.6881	1.0000	1.0000
PERMARKT	99.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

BISERIAL CORR
SCALE-ITEM 99.0000 0.5141 0.6315 0.5739 0.6162 0.6281 0.9323

116

SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM

ORDER FROM MCGRAW-HILL: SPSS, 2ND ED. (PRINCIPAL TEXT)
SPSS PRIMER (BRIEF INTRO TO SPSS)
SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)

ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS
SPSS POCKET GUIDE, RELEASE 8
KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION..
WORKSPACE 17920 BYTES
TRANSPACE 2560 BYTES

ALLOWS FOR..
25 TRANSFORMATIONS
102 RECODE VALUES + LAG VARIABLES
412 IF/COMPUTE OPERATIONS

```
  1 GET FILE      RSOC
```

FILE RSOC HAS 409 VARIABLES

THE SUBFILES ARE..

NAME	NO OF CASES
WEST	24
SBEI	11
CNTR	17
NRTH	19
JEND	9

CPU TIME REQUIRED.. 0.56 SECONDS

```
  2 SELECT IF     (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207604))
  3 COMPUTE       PERMARKT=0
  4 IF            (ID EQ 104006)PERMARKT=1
  5 IF            (ID EQ 206004)PERMARKT=1
  6 IF            (ID EQ 208003)PERMARKT=1
  7 IF            (ID EQ 309008)PERMARKT=1
  8 *RECODE       V002(0 THRU HIGHEST=1)
  9 *RECODE       V302(1 THRU HIGHEST=1)
 10 *RECODE       V303(1 THRU HIGHEST=1)
 11 *RECODE       V321(1 THRU HIGHEST=1)
 12 WRITE CASES   (F6.0,4X,7(F2.0))ID  PERMARKT  V099  V249  V321  V303  V302  V002
```

```
76 CASES WRITTEN ON LOGICAL UNIT # 3
 0 (UNWEIGHTED) CASES WERE DROPPED DUE TO MISSING VALUES

TRANSPACE REQUIRED..      1000 BYTES
   16 TRANSFORMATIONS
    8 RECODE VALUES + LAG VARIABLES
   38 IF/COMPUTE OPERATIONS

CPU TIME REQUIRED..    0.32 SECONDS

          13 FINISH          (SPSS GENERATED)

   NORMAL END OF JOB.
      13 CONTROL CARDS WERE PROCESSED.
       0 ERRORS WERE DETECTED.
```

101001	0	0	0	0	0	1	1	
101002	0	0	1	1	1	1	1	
101003	0	0	0	0	0	0	1	
101004	0	1	1	1	0	1	1	
101005	0	0	0	0	1	1	1	
101006	0	0	0	0	0	1	1	
101007	0	0	0	0	0	1	1	
101008	0	0	0	0	0	1	1	
101009	0	0	1	0	1	1	1	-
102003	0	0	1	1	1	1	1	
102004	0	1	1	1	1	1	1	
102005	0	0	1	1	1	1	1	
102006	0	0	0	0	0	0	1	
102007	0	0	0	0	0	0	1	
102008	0	0	0	0	0	1	1	
102009	0	1	1	1	1	1	1	
102010	0	0	0	0	0	0	1	
102011	0	0	0	0	0	0	1	
102012	0	0	0	0	0	0	1	
102013	0	0	0	0	0	1	1	
102014	0	0	0	0	0	1	1	
102015	0	0	1	1	1	1	1	
103001	0	0	1	0	0	1	1	-
103002	0	0	1	1	0	1	1	-
103003	0	0	0	0	1	1	1	
103004	0	0	0	1	1	1	1	
103005	0	0	0	1	0	0	1	-
103006	0	0	0	1	1	1	1	
103007	0	0	0	1	0	1	1	-
103008	0	0	0	0	0	0	1	
103009	0	0	0	0	0	0	1	
103010	0	0	0	1	0	1	1	-
104001	0	0	1	0	0	0	1	-
104002	0	0	0	0	0	1	1	
104003	0	0	1	0	1	1	1	-
104004	0	0	0	1	0	1	1	-
104005	0	0	0	0	0	0	1	
104006	1	1	1	1	1	1	1	
104007	0	0	0	0	0	1	1	
105001	0	0	0	0	0	0	1	
105002	0	1	0	0	0	1	1	— 2, 3 or 6?
105003	0	0	0	0	1	0	1	-
105004	0	1	0	0	0	1	1	—
105005	0	0	1	0	0	1	1	-
105006	0	0	0	0	1	1	1	
105007	0	0	1	1	1	1	1	
105008	0	0	0	0	1	0	1	-
105009	0	1	0	0	0	0	1	—
105010	0	0	0	0	0	0	1	
206001	0	0	1	0	1	1	1	-
206002	0	1	0	1	0	0	1	-
206003	0	0	1	1	1	0	1	-
206004	1	1	1	1	1	1	1	
206005	0	0	0	0	0	0	1	
206006	0	0	0	1	0	0	1	-
207001	0	0	0	0	0	0	1	
207002	0	0	0	1	1	1	1	
207003	0	0	1	0	1	1	1	-
207005	0	0	0	1	1	1	1	
207006	0	0	1	[illegible]	0	0	1	-
207007	0	0	1	[illegible]	0	0	1	
207008	0	0	1	1	1	1	1	
208001	[illegible]	0	1	[illegible]	[illegible]	[illegible]	[illegible]	
208002	0	0	0	0	1	1	1	

208003	1	1	1	1	1	1	1	
208004	0	0	0	0	0	0	1	
209005	0	0	0	1	1	0	1	-
309001	0	1	0	1	1	1	1	-
309002	0	0	1	0	1	1	1	-
309003	0	1	0	1	0	0	1	-
309004	0	0	0	1	1	1	1	
309005	0	0	0	0	1	0	1	-
309006	0	1	1	1	0	1	1	-
309007	0	0	0	1	1	1	1	
309008	1	1	1	1	1	1	1	
309009	0	0	0	1	1	1	1	

EXERCISE 6: CONSTRUCTING AN INDEX BASED ON THE Z-SCORE TECHNIQUE

Problem: How does one construct an index with variables whose distributions are dissimilar?

Readings: SPSS:187-188, Baseline: p. 56-65.

Short Answer: Decide on a concept such as "level of living." Identify reliable items by using the reliability technique. Then transform the distribution of each of these items by means of z-scoring, and add the values for the four items. The resulting index can be referred to as a "composite score," a "composite z-score," or one can simply use the substantive title, such as "level of living score."

Output: Pages A1-4; B1-6

Background: In previous exercises, you have seen some of the problems of selecting items in terms of concepts, you have looked at the reliability problem, and you have learned simple scores and Guttman scaling. Z-scoring is a third technique that is relevant to a set of items with dissimilar distributions. The effect of the z-score transformation is to set all the distributions to a mean of 0 and a standard deviation of 1. As it is often said, it is a technique to allow you to add "apples and oranges," but one should not overdo it by throwing in a cauliflower or an artichoke. Use the reliability program and common sense to screen out items that are not consistent with your concept. This technique has been widely used for "level of living scales," which are one measure of poverty.

Steps:

1. The first section of the output shows the reliability test for the four useable items from the section of the questionnaire dealing with housing and possessions. Note that these four items apply to the dryland or poor farmers only. The correlation matrix is favorable in that all of the correlations are positive, and the mean correlation is respectable, .42.

The correlation of each item with the whole, as shown in the middle column of the output, is also respectable, although not great. In particular, the correlation between the proportion of purchased doors on the houses of the dryland farmers is lower than the others (.35 as compared to .50.) As for the change in the alpha coefficient if the items are deleted, there does seem to be a great deal of variation there which puzzles me considering that the standardized item alpha is .74. The project decided to live with these results if for no other reason than we had little alternative.

2. Now that we have ascertained as best we can that we are dealing with apples and oranges and not odd vegetables, the next step is to standardize the distributions. The formula for z-scores is as follows:

$$Z = \frac{X_i - \bar{X}}{S_x} \quad \text{or} \quad \frac{\text{value for one unit - mean for all units}}{\text{standard deviation for X}}$$

Notice how you compute the formula, as shown on page B1, line 8. All you do is subtract the mean of the distribution from the value for a given subdistrict and divide by the standard deviation. Where do you get these numbers? The value for the percent of windows / is simply the label, DWIND, and that is enough for the computer to find the number. The other numbers are on page 2 of the reliability output, or you could turn back to the CONDESCRIPTIVE output for the same numbers. So it is simply a matter of typing up the control cards correctly. Then, as shown in line 12, you compute the composite z-score here labeled ZHOUSE by adding up the four new variables. If you wanted to, you could divide by four except that it would not mean much in this case so we did not bother. Then, just to make sure it all came out right, you print out the first ten cases, as shown on page 3. At this point, if you really want to understand this stuff, you should calculate the first value by hand. That is, go

back and get the actual value for Case No. 1 and work out the z-score for windows. Then, add up the four z-scores there on page 3 to see that they actually add up to the z-score that is listed under ZHOUSE.

3. Continuing along with the checking operation, have the computer do CONDESCRIPTIVES. The results appear on page B4. Note that all the means are 0, and the standard deviations are 1. That is the effect of z-scoring. There are still some problems with kurtosis and skewness, but these will be dealt with in a subsequent exercise on "transformations."

One last thing: Study the text and table (3.1) on pp. 56ff of <u>Baseline</u>. That is how it all came out.

```
*****************SYSMSG   ************RM3.PR1 ***FMT <00>* *****198 LINES *** 12:17.36AM  24 NOV 81  WARREN                SUSAN BENTLEY
B | STEP NAME                    SPSS     |
B | CPU SECONDS                     1.27 |
B | CPU AU'S (CPU*5.00)             6.35 |
B |TOTAL I/O TIME, SECS            19.85 |
B | REGION USED                     148K |
B ***************************************************************************************************************
B * BEST CLASS FOR JOB- A *|          USED     LIMIT            MIN:SEC                                  REGION          TOTAL    |
B *                       *|                                     USED   LIMIT  $CHARGED                                    EXECUTION |
B * AMOUNT SAVED BY       *| LINES       101      3000                                                  MAX USED    140K     CHARGE   |
B * RUNNING IN THIS CLASS *| CARDS         0       500   CPU     :01     2:00      $.12                LIMIT       160K             |
B * WOULD BE:        $.01 *| FILING        0       500   I/O     :20     5:00      $.38                CHARGE      $.52        $1.02 |
B ***************************************************************************************************************

0 JOB 2030-EM213     FROM RM3.R01   15:26.05 MONDAY NOVEMBER 23, 1981
0                     //EM213    JOB Y8H$$$$,'SUSAN BENTLEY        ',PRTY=4                                  DLVR=WRN
C                     /*LIM REGION=160K
C                     // EXEC SPSS
C                     //FT03F001  DD  UNIT=SYSDA,DISP=SHR,DSN=Y8H.RSOC
C                       INPUT STREAM DATASET OF 10 CARDS
C                     //

B *-----------------------------------ACCOUNTING CONTROL RECORD (ACR) DISPLAY-----------------------------------*
B | Y8H    E S 213 RM                ACCOUNT NO.:  U9G4F7AA99     STATUS:   ACTIVE       ACCT BALANCE:      $168.89   |
C | DFLT RETPD: SYS        DSACCESS: SYS  DASD+CF PER DAY CHG:   $.13  DASD+CF TOT CHG:       $19.66                |
C | HASP TOTAL CHG:     $106.51                HASP CHG LIMIT:     $300.00                HASP BALANCE:      $193.49   |
B *-------------------------------------------------------------------------------------------------------------*
C //EM213    JOB Y8H$$$$,'SUSAN BENTLEY        ',PRTY=4                                  DLVR=WRN
C // EXEC SPSS
C XXSPSS PROC SPACE=20K,TYPE1=800,TYPE2=2012,TYPE3=TRK,                                  00000010
C XX       MAIN1=250,MAIN2=750,MAIN3=80,MAIN4=2250,                                      00000020
C XX       SEC1=50,SEC2=190,LIB='SYS2.SPSS.LOAD',VERSION=H,                              00000030
C XX       ARCHIVE='DUMMY,'                                                              00000040
C ***                                                                                    00000040
C *** CURRENTLY ACCESSING THE INSTALLED VERSION OF SPSS 8.1 (05/20/80)*** 00000050
C *** PROBLEMS SHOULD BE REPORTED TO USER SERVICES AT 607-256-4981.   *** 00000060
C ***                                                                 *** 00000070
C *** PROCEDURE   DEFAULTS  ALLOW  SUFFICIENT  SCRATCH  SPACE   FOR   *** 00000080
C *** PROCESSING UP TO  1500 CASES WITH 250 VARIABLES.  OR MORE CASES *** 00000090
C *** WITH FEWER VARIABLES.  IF  MORE SPACE  IS  NEEDED, USERS  MAY   *** 00000100
C *** INCREASE IT ON THE FT01, FT02, &  SCRTWK JCL CARDS BY USING THE *** 00000110
C *** MAIN1, MAIN2, MAIN3, SEC1, SEC2, & TYPE3 SUBSTITUTION PARAMETERS*** 00000120
C ***                                                                 *** 00000130
C *** IF 'GET ARCHIVE', 'SAVE ARCHIVE', OR 'MERGE FILES' IS TO BE USED*** 00000140
C *** THEN CODE ARCHIVE=  ON THE EXEC CARD IN ORDER TO ACTIVATE THE   *** 00000150
C *** NECESSARY SCRATCH SPACE.                                        *** 00000160
C ***                                                                 *** 00000170
C *** FOR THE MAXI VERSION OF SPSS CODE THE FOLLOWING ON THE EXEC CARD*** 00000180
C ***                        // EXEC SPSS,VERSION=M                   *** 00000190
C XXSPSS     EXEC PGM=SPSS&VERSION,PARM=%SPACE                        *** 00000200
C IEF653I SUBSTITUTION JCL - PGM=SPSSH,PARM=20K                                          00000210
C XXSTEPLIB  DD DSN=&LIB,DISP=SHR                                                        00000220
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSS.LOAD,DISP=SHR
C XXFT01F001 DD UNIT=SYSSC,SPACE=(&TYPE1,(&MAIN1,&SEC1)),DSN=&&LABELS,                   00000230
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(800,(250,50)),DSN=&LABELS,
C XX            DISP=(,DELETE)                                                           00000240
C XXFT02F001 DD UNIT=SYSSC,SPACE=(&TYPE2,(&MAIN2,&SEC2)),DSN=&&DATASPSS,                 00000250
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(2012,(750,190)),DSN=&DATASPSS,
C XX            DISP=(,DELETE)
C XXFT05F001 DD DUMMY=SYSIN                                                              00000260
C XXFT06F001 DD SYSOUT=A                                                                 00000270
C XXFT10F001 DD DSN=SYS2.SPSS&VERSION..STAT,DISP=SHR                                     00000280
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSSH.STAT,DISP=SHR                                00000290
```

A

SPSS BATCH SYSTEM

SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

ORDER FROM MCGRAW-HILL: CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM
SPSS, 2ND ED. (PRINCIPAL TEXT)
SPSS PRIMER (BRIEF INTRO TO SPSS)
SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)

ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS
SPSS POCKET GUIDE, RELEASE 8
KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION.. ALLOWS FOR..
WORKSPACE 17920 BYTES 25 TRANSFORMATIONS
TRANSPACE 2560 BYTES 102 RECODE VALUES + LAG VARIABLES
412 IF/COMPUTE OPERATIONS

```
1 GET FILE      RSOC
```

FILE RSOC HAS 409 VARIABLES

THE SUBFILES ARE..

NAME	NO OF CASES
WEST	24
SBET	11
CNTR	17
NRTH	19
JEND	9

CPU TIME REQUIRED.. 0.59 SECONDS

```
2 COMPUTE          DWIND=(V215/V211)*100
3 COMPUTE          DDOOR=(V216/V211)*100
4 COMPUTE          DSTOVE=(V217/V211)*100
5 COMPUTE          DSEW=(V218/V211)*100
6 ASSIGN MISSING   DWIND TO DSEW(-1)
7 SELECT IF        (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
8 RELIABILITY      VARIABLES=DWIND TO DSEW/
9                  SCALE(DRYLAND)=DWIND TO DSEW/
10 STATISTICS      1,3,8,9
```

****** METHOD 2 (COVARIANCE MATRIX) WILL BE USED FOR THIS ANALYSIS ******

****** 336 BYTES OF SPACE REQUIRED FOR RELIABILITY ******

2

FILE RSOC (CREATION DATE = 11/04/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * * * * R E L I A B I L I T Y A N A L Y S I S F O R S C A L E (D R Y L A N D)* * * * * * * * *

1. DWIND
2. DDOOR
3. DSTOVE
4. DSEW

		MEANS	STD DEV	CASES
1.	DWIND	22.10971	27.54676	74.0
2.	DDOOR	49.22708	36.08117	74.0
3.	DSTOVE	8.27379	15.88351	74.0
4.	DSEW	0.73467	1.25178	74.0

CORRELATION MATRIX

	DWIND	DDOOR	DSTOVE	DSEW
DWIND	1.00000			
DDOOR	0.36419	1.00000		
DSTOVE	0.57066	0.23486	1.00000	
DSEW	0.43433	0.23765	0.68421	1.00000

OF CASES = 74.0

INTER-ITEM CORRELATIONS	MEAN	MINIMUM	MAXIMUM	RANGE	MAX/MIN	VARIANCE
	0.42098	0.23486	0.68421	0.44934	2.91321	0.02970

ITEM-TOTAL STATISTICS

	SCALE MEAN IF ITEM DELETED	SCALE VARIANCE IF ITEM DELETED	CORRECTED ITEM-TOTAL CORRELATION	SQUARED MULTIPLE CORRELATION	ALPHA IF ITEM DELETED
DWIND	57.23554	1873.57789	0.52554	0.38292	0.25449
DDOOR	31.11317	1569.20676	0.35493	0.14098	0.53199
DSTOVE	71.07146	2837.60596	0.47026	0.56061	0.40987
DSEW	78.61036	3805.47188	0.50412	0.47543	0.58830

126

* * * * * * * * * R E L I A B I L I T Y A N A L Y S I S F O R S C A L E (D R Y L A N D) * * * * * * * * * *

RELIABILITY COEFFICIENTS 4 ITEMS

ALPHA = 0.53912 STANDARDIZED ITEM ALPHA = 0.74413

```
TRANSPACE REQUIRED..        500 BYTES
      5 TRANSFORMATIONS
      0 RECODE VALUES + LAG VARIABLES
     32 IF/COMPUTE OPERATIONS

CPU TIME REQUIRED..     0.37 SECONDS


             11 FINISH          (SPSS GENERATED)

        NORMAL END OF JOB.
           11 CONTROL CARDS WERE PROCESSED.
            0 ERRORS WERE DETECTED.
```

```
C XXFT06F001 DD STSOUT=A
C XXFT10F001 DD DSN=SYS2.SPSS&VERSION..STAT,DISP=SHR
```

```
*****************SYSMSG    ************RM3.PR1 ***FMT <00>* ****198 LINES *** 12:27.56PM   5 DEC 81   WARREN          SUSAN BENTLEY
B | STEP NAME                    SPSS    |
B | CPU SECONDS                     1.25 |
B | CPU AU'S (CPU*5.00)             6.25 |
B | TOTAL I/O TIME, SECS           19.76 |
B | REGION USED                     140K |
B ***************************************
B * BEST CLASS FOR JOB- A *|
B *                       *|
B * AMOUNT SAVED BY       *|
B * RUNNING IN THIS CLASS *|
B * WOULD BE:       $.02  *|
B ***************************
```

	USED	LIMIT		MIN:SEC USED	MIN:SEC LIMIT	$CHARGED
LINES	136	3000				
CARDS	0	500	CPU	:01	2:00	$.12
FILING	0	500	I/O	:20	5:00	$.38

REGION		TOTAL EXECUTION CHARGE
MAX USED	148K	
LIMIT	160K	
CHARGE	$.51	$1.01

```
B JOB 7543-EM837     FROM RM3.RD1   13:42.07 FRIDAY DECEMBER 4, 1981
B                    //EM837    JOB Y8H$$$$,'SUSAN BENTLEY        ',PRTY=4                    DLVR=WRN
C                    /*LIM REGION=160K
C                    // EXEC SPSS
C                    //FT03F001  DD  UNIT=SYSDA,DISP=SHR,DSN=Y8H.RSOC
C                      INPUT STREAM DATASET OF 17 CARDS
C                    //


B *-----------------------------------ACCOUNTING CONTROL RECORD (ACR) DISPLAY-----------------------------------*
B | Y8H   R S 213 NW                ACCOUNT NO.:  U964F7AA99    STATUS:   ACTIVE      ACCT BALANCE:     $150.04   |
C | DFLT RETPD: SYS          DSACCESS: SYS  DASD+CF PER DAY CHG:   $.13  DASD+CF TOT CHG:      $23.07          |
C | HASP TOTAL CHG:     $120.35                HASP CHG LIMIT:     $300.00                HASP BALANCE:     $179.65   |
B *-----------------------------------------------------------------------------------------------------------*
C //EM837    JOB Y8H$$$$,'SUSAN BENTLEY        ',PRTY=4
C // EXEC SPSS                                                                          DLVR=WRN
C XXSPSS PROC SPACE=20K,TYPE1=800,TYPE2=2012,TYPE3=TRK,                                 00000010
C XX       MAIN1=250,MAIN2=750,MAIN3=80,MAIN4=2250,                                     00000020
C XX       SEC1=50,SEC2=190,LIB='SYS2.SPSS.LOAD',VERSION=H,                             00000030
C ***                                                                                   00000040
C *** CURRENTLY ACCESSING THE INSTALLED VERSION OF SPSS 8.1 (05/20/80)*** 00000050
C *** PROBLEMS SHOULD BE REPORTED TO USER SERVICES AT 607-256-4981.   *** 00000060
C ***                                                                   *** 00000070
C *** PROCEDURE   DEFAULTS  ALLOW   SUFFICIENT   SCRATCH   SPACE   FOR *** 00000080
C *** PROCESSING UP TO  1500 CASES WITH 250 VARIABLES.  OR MORE CASES *** 00000090
C *** WITH  FEWER VARIABLES.   IF  MORE SPACE  IS  NEEDED, USERS  MAY *** 00000100
C *** INCREASE IT ON THE FT01, FT02, &  SORTWK JCL CARDS BY USING THE *** 00000110
C *** MAIN1, MAIN2, MAIN3, SEC1, SEC2, & TYPE3 SUBSTITUTION PARAMETERS*** 00000120
C ***                                                                   *** 00000130
C *** IF 'GET ARCHIVE', 'SAVE ARCHIVE', OR 'MERGE FILES' IS TO BE USED*** 00000140
C *** THEN CODE ARCHIVE=  ON THE EXEC CARD IN ORDER TO ACTIVATE THE   *** 00000150
C *** NECESSARY SCRATCH SPACE.                                          *** 00000160
C ***                                                                   *** 00000170
C *** FOR THE MAXI VERSION OF SPSS CODE THE FOLLOWING ON THE EXEC CARD*** 00000180
C ***                                                                   *** 00000190
C XXSPSS     EXEC PGM=SPSS&VERSION,PARM='SPACE                          *** 00000200
C IEF653I SUBSTITUTION JCL - PGM=SPSSH,PARM=20K                                 00000210
C XXSTEPLIB  DD DSN=&LIB,DISP=SHR
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSS.LOAD,DISP=SHR                        00000220
C XXFT01F001 DD UNIT=SYSSC,SPACE=(&TYPE1,(&MAIN1,&SEC1)),DSN=&&LABELS,          00000230
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(800,(250,50)),DSN=&LABELS,
C XX            DISP=(,DELETE)
C XXFT02F001 DD UNIT=SYSSC,SPACE=(&TYPE2,(&MAIN2,&SEC2)),DSN=&&DATASPSS,        00000240
C IEF653I SUBSTITUTION JCL - UNIT=SYSSC,SPACE=(2012,(750,190)),DSN=&DATASPSS,   00000250
C XX            DISP=(,DELETE)
C XXFT05F001 DD DDNAME=SYSIN                                                    00000260
C XXFT06F001 DD SYSOUT=A                                                        00000270
C XXFT10F001 DD DSN=SYS2.SPSS&VERSION.,STAT,DISP=SHR                            00000280
C IEF653I SUBSTITUTION JCL - DSN=SYS2.SPSSH.STAT,DISP=SHR                       00000290
```

B

```
*********SPSS     FT06F001 ************RM3.PR1 ***FMT <00>* *****136 LINES *** 12:29.37PM   5 DEC 81  WARREN          SUSAN BENTLEY
SPSS BATCH SYSTEM
                                                                                  12/05/81        PAGE    1
SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

                          CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM
ORDER FRO MCGRAW-HILL:  SPSS, 2ND ED. (PRINCIPAL TEXT)        ORDER FROM SPSS INC.:  SPSS STATISTICAL ALGORITHMS
                        SPSS PRIMER (BRIEF INTRO TO SPSS)                            SPSS POCKET GUIDE, RELEASE 8
                        SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)                  KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION..      ALLOWS FOR..     25 TRANSFORMATIONS
WORKSPACE      17920 BYTES                     102 RECODE VALUES + LAG VARIABLES
TRANSPACE       2560 BYTES                     412 IF/COMPUTE OPERATIONS

          1 GET FILE      RSOC


                    FILE RSOC     HAS 409 VARIABLES

                    THE SUBFILES ARE..

                              NO OF
                    NAME      CASES

                    WEST        24
                    SRET        11
                    CNTR        17
                    NRTH        19
                    JENO         9

CPU TIME REQUIRED..     0.55 SECONDS


          2 SELECT IF       (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
          3 COMPUTE         DWIND=(V215/V211)*100
          4 COMPUTE         DDOOR=(V216/V211)*100
          5 COMPUTE         DSTOVE=(V217/V211)*100
          6 COMPUTE         DSEW=(V218/V211)*100
          7 ASSIGN MISSING  DWIND TO DSEW(-1)
          8 COMPUTE         ZWIND=(DWIND-22.110)/27.547
          9 COMPUTE         ZDOOR=(DDOOR-48.227)/36.081
         10 COMPUTE         ZSTOVE=(DSTOVE-8.274)/15.884
         11 COMPUTE         ZSEW=(DSEW-0.735)/1.252
         12 COMPUTE         ZHOUSE=ZWIND+ZDOOR+ZSTOVE+ZSEW
         13 ASSIGN MISSING  ZWIND TO ZHOUSE(99999)
         14 PRINT FORMATS   ZWIND TO ZHOUSE(5)
         15 LIST CASES      CASES=10/VARIABLES=ID,ZWIND TO ZHOUSE
         16 CONDESCRIPTIVE  ZWIND TO ZHOUSE

***** GIVEN WORKSPACE ALLOWS FOR  358 VARIABLES FOR CONDESCRIPTIVE PROBLEM *****
```

17 Statistics All

130

CASE-N	ID	ZWIND	ZDOOR	ZSTOVE	ZSEW	ZHOUSE
1	101001.	-0.14920	0.04914	-0.50516	-0.58706	-1.19228
2	101002.	-0.80263	1.43491	-0.52090	-0.58706	-0.47568
3	101003.	-0.69093	-1.12344	0.44766	0.88750	-0.47920
4	101004.	0.75315	-0.14883	-0.43096	0.32576	0.49913
5	101005.	-0.51105	-0.22356	0.74328	-0.42667	-0.41800
6	101006.	-0.31861	-0.96709	-0.31105	1.18788	-0.40887
7	101007.	2.82753	1.43491	0.73823	-0.50706	4.41361
8	101008.	0.04441	0.60545	-0.52090	0.47790	0.60486
9	101009.	2.10150	1.43491	1.24138	0.69089	5.46918
10	102003.	1.27175	0.24711	-0.16115	0.55397	1.91167

FILE RSOC (CREATION DATE = 11/04/81)
SUBFILE WEST SBEI CNTR NRTH JEND

VARIABLE ZKINP

MEAN	-0.000	STD ERROR	0.116	STD DEV	1.000
VARIANCE	1.000	KURTOSIS	1.668	SKEWNESS	1.569
RANGE	3.630	MINIMUM	-0.803	MAXIMUM	2.828
SUM	-0.001				

VALID OBSERVATIONS - 74 MISSING OBSERVATIONS - 2

- -

VARIABLE ZLCOR

MEAN	0.000	STD ERROR	0.116	STD DEV	1.000
VARIANCE	1.000	KURTOSIS	-1.411	SKEWNESS	0.230
RANGE	2.772	MINIMUM	-1.337	MAXIMUM	1.435
SUM	0.000				

VALID OBSERVATIONS - 74 MISSING OBSERVATIONS - 2

- -

VARIABLE ZSTOVE

MEAN	-0.000	STD ERROR	0.116	STD DEV	1.000
VARIANCE	1.000	KURTOSIS	13.471	SKEWNESS	3.449
RANGE	5.771	MINIMUM	-0.521	MAXIMUM	5.250
SUM	-0.001				

VALID OBSERVATIONS - 74 MISSING OBSERVATIONS - 2

- -

VARIABLE ZSEW

MEAN	-0.000	STD ERROR	0.116	STD DEV	1.000
VARIANCE	1.000	KURTOSIS	5.741	SKEWNESS	2.381
RANGE	4.564	MINIMUM	-0.587	MAXIMUM	3.977
SUM	-0.019				

VALID OBSERVATIONS - 74 MISSING OBSERVATIONS - 2

- -

SPSS BATCH SYSTEM

FILE RSOC (CREATION DATE = 11/04/81) 12/05/81 PAGE 5
SUBFILE WEST SBEI CNTR NRTH JEND

VARIABLE ZFCUSE

MEAN	-0.000	STD ERROR	0.350	STD DEV	3.008
VARIANCE	9.051	KURTOSIS	5.108	SKEWNESS	2.014
RANGE	16.166	MINIMUM	-3.247	MAXIMUM	12.919
SUM	-0.021				

VALID OBSERVATIONS - 74 MISSING OBSERVATIONS - 2

Best Available Document

```
TRANSPACE REQUIRED..     1000 BYTES
     10 TRANSFORMATIONS
      6 RECODE VALUES + LAG VARIABLES
     53 IF/COMPUTE OPERATIONS

CPU TIME REQUIRED..    0.39 SECONDS


          18 FINISH          (SPSS GENERATED)

    NORMAL END OF JOB.
        18 CONTROL CARDS WERE PROCESSED.
         0 ERRORS WERE DETECTED.
```

EXERCISE 7: CROSS-TABULATION

Problem: What is the simplest way to explore relationships with categorical data?

Brief Answer: Use cross-tabulation. That is, run "contingency tables," compute percentages and possibly some coefficient of relationship, and present the data in this form.

Reading: SPSS: Ch. 16.

Output: A1-14; B1-4.

Commentary: I have heard skilled analysts remark that nothing gives them such a sense of security about data analysis than a nice strong cross-tabulation. There is good reason for this sentiment because cross-tabulation is simple and straightforward. It does not involve all the pitfalls of correlational or regression analysis. Other considerations are that administrators are more likely to be able to read tables based on cross-tabulation and, if you are in the field, you can do the analysis by hand.

Cross-tabulation tables are also quite sensitive to non-linear relationships, which are often difficult to detect in correlational or regression analysis. Indeed, it has been said that one reason why the early quantifiers in sociology latched onto cross-tabulation instead of regression, which they certainly understood at the time, was that cross-tabulation allowed them to show relationships in a more precise and clear way. The fact that the computation of regressions was onerous without computers was a contributing factor along with the appropriateness of cross-tabulation for their sample surveys of political attitudes and other personal attributes. Cross-tabs have and will continue to have a place in sociological analysis. This is particularly so now that "log linear analysis" has been introduced, which permits one to do everything with cross-tabs that we now do with regression analysis.

What follows is essentially a lot of examples of the different ways that cross-tabs can be used. All the tables are variations on a theme, but it is important to grasp the theme.

Steps:

1. Page 1 shows the control cards, and you see that we recomputed the Ecoscore because we did not save it in the file. Under normal research conditions the score would have been saved, so you could have simply called out the Ecoscore. On line 4, the Ecoscore had to be recoded into a dichotomy in preparation for cross-tabulation. This operation "loses" information because now we no longer have a continuous variable. Similarly, the other variables were recoded. Line 8 shows the command for CROSSTABS, but note that the "independent variable," which in this case is V090, goes last. That is important.

2. Page 2 shows a simple cross-tabulation of the importance of commercial grasses by the presence of a water peddler. Commercial grasses is used here as an index of the ecology of the area because general observation

136

indicates that these grasses only occur in the very dry areas, and it is known that the women of the sedentary herding families collect them and sell them to a factory in the town of Kasserine where they are made into a high-grade paper. So the cross-tabulation between this variable and the presence of a water peddler is a way to validate this perception.

The general rule for cross-tabulation is: calculate percentages in the direction of presumed causality and read across. Or, if you have the "independent variable" at the top of the table as is true in this case, then percentage down and read across. That means that you should pay attention to the column of percentages, which is the third number in each cell. Furthermore, inasmuch as the dependent variable is dichotomous, you will only have to look at one line, which is the "present" line signified by the code "1." The relevant numbers are underlined, and you see that the percentage of subdistricts with at least one person selling water increases as commercial crops are rated to be very important. One would interpret this table simply as "there is a moderately strong relationship between importance of commerical crop and the presence of commercial water distribution." If you were publishing this table, it would only be necessary to publish that row of numbers together with the N's, which in this case are 40, 17, and 19. There is no need to present the other row of numbers, that is, where there is no commerical water distribution, because this is a dichotomy. Finally, when you construct titles for tables, a rule of thumb is to mention the dependent variable first. Thus, the title for this table might be: Relationship of commercial water distribution to the informant's estimate of the importance of commercial grasses in the subdistrict.

3. Page 3 shows a cross-tabulation between the importance of commercial grasses and the amount of land sold. Note that this variable has been recoded

so that "1" refers to 1-12 hectares "2" refers to 15-46 hectares and "3" refers to 50-400 hectares. How do you decide these cutting points? The basic rule is to divide the sample into more or less equal segments. A complementary rule is to pay attention to any sociologically relevant cutting points. If, for example, you knew that it takes about 20 hectares of land for subsistence in this area, then you might use that cut-off point. The first category used in the table is probably slightly below the subsistence level.

As before, you pay attention to the column percentages and read across. Thus, the numbers are 37.5, 29.4, and 26.3. There is a slight trend there, but it is not very strong. Next, you look at the "3" row, and you see that the numbers are not linear: 25.0, 52.9, and 47.4. There is a relationship there, but it holds only for areas where grasses are not important versus those districts where it has some or much importance. Thus, if you were using this table in a report, you would have to collapse categories 1 and 2 of V090. Try this, and recalculate the percentages. Notice that the first number in each cell is the actual number of cases.

If you wanted to, you could compute the correlation coefficient for this table. Some of these are computed in the end of this output in order to give you a feel for the relationship between percentage differences (which is a kind of measure of association) and an actual correlation coefficient. My guess is that the Pearson's R for this table is about .25. But you can run it yourself and check me.

4. Starting page 4 is a set of related dependent variables, and the point of these is to show you how separate cross-tabulations can be combined into one large table that is more compact and informative. The last page of this handout shows the format you should use. As you see, the relevant figures for the first table are already listed on the format page. Add the appropriate

figures from pages 5-9. As you can see, only V122 and V129 are clearly related to the importance of commercial grasses. Note that both of these pertain to water. On the one hand, the arid lands are vulnerable to heavy rains while, on the other, the irrigation leads to contamination of the water, probably by salinization. It may be, then, that you should rewrite your table so as to include only these two variables. Leave the others out or, alternatively, group them together below and only talk about the two water items. Either way, you can interpret the table in a more dramatic way now that you have both water items side by side. And, needless to say, the information is much more compact. Indeed, one way you can spot an untrained person is that he or she does not organize and edit the output from cross-tabulations.

5. The cross-tabulation on page 10 is somewhat puzzling. The Ecoscore combines precisely the variables that we have been talking about, only two of which had a strong relationship. It appears, however, that when the Ecoscore is dichotomized, it shows a moderately strong relationship with importance of commercial grasses. But this relationship is being "carried" by only two of the items in the score. Now you see why cross-tabulations are so revealing. A simple correlation analysis, as shown on page 14, gives a relationship of .31, and that is true as far as it goes; but only by analyzing each attribute can you tell what is really happening. However, the score of ecological problems is still valid and useful in a general way. What you have to realize is that, depending on the independent variable, different items in the score will account for the relationship. That is not the best kind of score to have; but in the case of ecological problems, but which tend to vary from place to place, perhaps that is the best that is possible. By the way, notice that on page 10 the percentage difference between the lowest and highest percent is 36 percent. That seems like a large gap and it is, but when it is converted to a correlation, the relationship is "only" .31. In other words, it takes a pretty big percentage difference to show up as a strong correlation.

6. Page 11 shows what seems like a simple cross-tabulation, but it is complicated by the fact that both variables had to be recoded. As you see, there is a relationship there, and it is possible to show it in cross-tabular form, but the question is why should you when both variables are continuous? The correlation between these two continuous variables, as shown on page 14, is -.21.

7. Page 12 shows a relationship between the number of retail shops (dichotomized) which, as the correlation coefficient on page 14 indicates, is essentially a zero relationship. The correlation is actually -.04. What is of note here is the slight curvilinear relationship. The relevant percentages go: 51.9, 42.9, and 47.6. In other words, there is a "dip" in the middle. That drop is not enough to take seriously, but if the middle number were something like .30, you might very well make an interpretation that depends on a curvilinear relationship. For example, you might argue that subdistricts with few retail shops are probably in low-density arid subdistricts where there are a lot of ecological problems. On the other hand, the subdistricts with a lot of retail shops have large populations where there are different ecological problems, such as cutting down all the trees for firewood with subsequent erosion. The point is that a zero correlation may hide a real relationship, and you can only find it by cross-tabulation.

8. Simple cross-tabulations are useful as far as they go, but it is usually necessary to introduce a control factor in order to clarify the relationships. An example of such a "third" factor is shown in B1-4. The dependent variable is Ecoscore; but, of course, the score had to be dichotomized so that there would be sufficient cases in each of the cells. The independent variable is importance of commercial grasses, which we are taking as a proxy for extreme aridity. Then, on the basis of some theory and general observation, we introduce the presence or absence of a mosque as a control. Essentially then, we need two cross-tabulations, one when a mosque is present and the

other when it is absent, and that is precisely what the computer does for you. It also calculates a grab bag of measures of association for each of the cross-tabulations, but for the moment we concentrate on the percentages.

The relationships that I had hypothesized and which I wanted to demonstrate are shown in the table on page B4. Believe it or not, this is the kind of summary of all this computer output that would go into a report and probably into a published paper. However, the output was not exactly in this form, so I had to rearrange the figures. Make sure that you understand where my figures came from; and as an exercise, write another table with the role of mosque and commercial grasses reversed. That is, show the relationship of ecological problems to presence of a mosque, controlling on the importance of commercial grasses.

The verbal description of the table would go as follows: Subdistricts where commercial grasses are important have many ecological problems, and this condition is especially pronounced if there is no mosque in the subdistrict. Or, somewhat more tersely: Importance of commercial grasses and the absence of a mosque are independently related to many ecological problems. Now, if you want to go on and give a more theoretical interpretation of these results, you might say the following: Inasmuch as the commercial grasses only grow in the very dry subdistricts, their presence is an indicator of a marginal area, and these tend to have more ecological problems by definition. However, it is also true that the sedentary herding families dwell in these subdistricts, and it may be that their animals, especially the goats, injure this vulnerable ecology. Remarkably, the presence of a mosque is associated with fewer problems. These mosques are located in the higher subdistricts where, along with commercial grasses, there are more trees and, in general, a more hardy ecology. The presence of a mosque reflects, at the minimum, a population

with a higher income, but it may also indicate a more organized and conservation-minded community.

9. Some comment is in order on the statistics that are available for cross-tabulation even though it would take a whole course to really explain them. First is the significance level as measured by chi-square. The table on page 2 is significant at the .03 level, which means that the result could occur only about three times out of a hundred by chance. In other words, this is probably a reliable result; or at any rate, it would occur by chance less than 5 percent of the time, and that is the arbitrary level of acceptability. However, as you recall, statistical significance is of no great importance in this study, inasmuch as we are not attempting to generalize to some universe. Some people do look at significance levels as a guide to how seriously they should take the result. Notice that the relationship on page 3 is not significant at the .05 level. On the contrary, its significance level is .44, which is almost equivalent to the flip of a coin.

What we are really interested in here is the measures of association, and there are two that are important. In the last row is our old friend Pearson's R, which in the top table is .40 and in the bottom table it is .14. Now if you believe that you should use some measure of association that is "non-parametric," that is, one that is not restricted by all the assumptions that hold for Pearson's R, then you might look at gamma, which is .69 in the top table and .37 in the lower one. Either way, the second table is a weaker result, but you can see that by comparing the percentages. In my table you see that the presence of a mosque does not make much difference when commercial grasses are not important. Presumably the ecology is generally good, so the level of wealth, organization, or whatever the mosques reflect, is not crucial. I might add that insofar as the Tunisians understood this and similar results, they were not happy because the present government is attempting to restrict

the role of Islam and they did not want to see any research that might be used as propaganda. However, I doubt whether any of the Mullahs bothered to read our report.

Table -- Relationship of Selected Ecological Problems to an Ecological Indicator: Importance of Commercial Grasses.

	Importance of commercial grasses		
	None	Important	Very important
Percent reporting:			
Deepening erosion	82	100	95
N =	(40)	(17)	(19)

Table -- Relationship of Selected Ecological Problems to an Ecological Indicator: Importance of Commercial Grasses.

	Importance of commercial grasses		
	None	Important	Very important
Percent reporting:			
Deepening erosion	82	100	95
N =	(40)	(17)	(19)

```
SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

ORDER FROM MCGRAW-HILL:   CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM
                          SPSS, 2ND ED. (PRINCIPAL TEXT)        ORDER FROM SPSS INC.:  SPSS STATISTICAL ALGORITHMS
                          SPSS PRIMER (BRIEF INTRO TO SPSS)                            SPSS POCKET GUIDE, RELEASE 8
                          SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)                  KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION..     ALLOWS FOR..     25 TRANSFORMATIONS
WORKSPACE     17920 BYTES                      102 RECODE VALUES + LAG VARIABLES
TRANSPACE      2560 BYTES                      412 IF/COMPUTE OPERATIONS

                1 GET FILE       RSOC


                          FILE RSOC     HAS 409 VARIABLES

                          THE SUBFILES ARE..

                                     NO OF
                          NAME       CASES

                          WEST         24
                          SBEI         11
                          CNTR         17
                          NRTH         19
                          JEND          9

CPU TIME REQUIRED..     0.55 SECONDS


                2 SELECT IF       (NOT (ID EQ 102001 OR 102002 OR 103011 OR 207004))
                3 COMPUTE         ECOSCORE=(V121+V122+V124+V125+V127+V129)
                4 RECODE          ECOSCORE(LO THRU 4=0)(5 THRU HI=1)
                5 RECODE          V075 (LO THRU 12=1) (15 THRU 46=2)(50 THRU 400=3)
                6 RECODE          V316(1 THRU HI=1)
                7 RECODE          V336(LO THRU 3=1)(4,5,6,7=2)(8 THRU HI=3)
                8 CROSSTABS       TABLES=V142,V075,V121,V122,V124,V125,V127,V129,ECOSCORE BY V090/
                9                 ECOSCORE BY V316/ECOSCORE BY V336

***** GIVEN WORKSPACE ALLOWS FOR    995 CELLS,    995 TABLES WITH     2 DIMENSIONS FOR CROSSTAB PROBLEM *****
```

146

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * C R O S S T A B U L A T I O N O F * * * * *
V142 SOMEONE SELLS WATER IN THE SUBDISTRICT BY V090 MAJOR CROP - COMMERCIAL GRASSES
* * * * * * PAGE 1 OF 1

V142	COUNT / ROW PCT / COL PCT / TOT PCT	V090 0.	1.	2.	ROW TOTAL
NO	0.	35 61.4 87.5 46.1	11 19.3 64.7 14.5	11 19.3 57.9 14.5	57 75.0
YES	1.	5 26.3 12.5 6.6	6 31.6 35.3 7.9	8 42.1 42.1 10.5	19 25.0
COLUMN TOTAL		40 52.6	17 22.4	19 25.0	76 100.0

147

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * * * * C R O S S T A B U L A T I O N O F * * * * * * * * *
V075 # HECTARES SOLD IN LAST 5 YEARS BY V090 MAJOR CROP - COMMERCIAL GRASSES
* PAGE 1 OF 1

V075	COUNT / ROW PCT / COL PCT / TOT PCT	V090 0.	1.	2.	ROW TOTAL
1.		15	5	5	25
		60.0	20.0	20.0	32.9
		37.5	29.4	26.3	
		19.7	6.6	6.6	
2.		15	3	5	23
		65.2	13.0	21.7	30.3
		37.5	17.6	26.3	
		19.7	3.9	6.6	
3.		10	9	9	28
		35.7	32.1	32.1	36.8
		25.0	52.9	47.4	
		13.2	11.8	11.8	
COLUMN TOTAL		40	17	19	76
		52.6	22.4	25.0	100.0

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * * * * * * * * * * * * * C R O S S T A B U L A T I O N O F * * * * * * * * * * * * * * * * * *
V121 DEEPENING EROSION OF RIVERS & GULLIES BY V090 MAJOR CROP - COMMERCIAL GRASSES
* PAGE 1 OF 1

V121	COUNT / ROW PCT / COL PCT / TOT PCT (V090)	0.	1.	2.	ROW TOTAL
NO	0.	7		1	8
		87.5	0.0	12.5	10.5
		17.5	0.0	5.3	
		9.2	0.0	1.3	
YES	1.	33	17	18	68
		48.5	25.0	26.5	89.5
		82.5	100.0	94.7	
		43.4	22.4	23.7	
	COLUMN TOTAL	40	17	19	76
		52.6	22.4	25.0	100.0

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * * * * * * * * * * * * * C R O S S T A B U L A T I O N O F * * * * * * * * * * * * * * * * * *
V122 ENLARGEMENT OF INUNDATED AREAS BY V090 MAJOR CROP - COMMERCIAL GRASSES
* PAGE 1 OF 1

V122	V090: COUNT / ROW PCT / COL PCT / TOT PCT	0.	1.	2.	ROW TOTAL
NO	0.	22	7	6	35
		62.9	20.0	17.1	46.7
		56.4	41.2	31.6	
		29.3	9.3	8.0	
YES	1.	17	10	13	40
		42.5	25.0	32.5	53.3
		43.6	58.8	68.4	
		22.7	13.3	17.3	
	COLUMN TOTAL	39	17	19	75
		52.0	22.7	25.3	100.0

NUMBER OF MISSING OBSERVATIONS = 1

153

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * * * * * * * * * * * * C R O S S T A B U L A T I O N O F * * * * * * * * * * * * * * * * * * *
V124 REMOTENESS FROM PASTURES BY V090 MAJOR CROP - COMMERCIAL GRASSES
* PAGE 1 OF 1

V124	COUNT / ROW PCT / COL PCT / TOT PCT	V090 0.	1.	2.	ROW TOTAL
NO	0.	14	6	6	26
		53.8	23.1	23.1	34.2
		35.0	35.3	31.6	
		18.4	7.9	7.9	
YES	1.	26	11	13	50
		52.0	22.0	26.0	65.8
		65.0	64.7	68.4	
		34.2	14.5	17.1	
	COLUMN TOTAL	40	17	19	76
		52.6	22.4	25.0	100.0

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * * * * * C R O S S T A B U L A T I O N O F * * * * * * * * * *
V125 DIFFICULTY IN FINDING FIREWOOD BY V090 MAJOR CROP - COMMERCIAL GRASSES
* PAGE 1 OF 1

V125	COUNT / ROW PCT / COL PCT / TOT PCT	V090 0.	1.	2.	ROW TOTAL
NO	0.	6	0	2	8
		75.0	0.0	25.0	10.5
		15.0	0.0	10.5	
		7.9	0.0	2.6	
YES	1.	34	17	17	68
		50.0	25.0	25.0	89.5
		85.0	100.0	89.5	
		44.7	22.4	22.4	
	COLUMN TOTAL	40	17	19	76
		52.6	22.4	25.0	100.0

* * * * * * * * * * * * * * * * * * C R O S S T A B U L A T I O N O F *
V127 REMOTENESS FROM WATER POINTS BY V090 MAJOR CROP - COMMERCIAL GRASSES
* PAGE 1 OF 1

V127	COUNT / ROW PCT / COL PCT / TOT PCT	V090 0.	1.	2.	ROW TOTAL
NO	0.	12	1	2	15
		80.0	6.7	13.3	19.7
		30.0	5.9	10.5	
		15.8	1.3	2.6	
YES	1.	28	16	17	61
		45.9	26.2	27.9	80.3
		70.0	94.1	89.5	
		36.8	21.1	22.4	
	COLUMN TOTAL	40	17	19	76
		52.6	22.4	25.0	100.0

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SDEI CNTR NRTH JEND

* * * * * * * * * * * * * * * * * * C R O S S T A B U L A T I O N O F * * * * * * * * * * * * * * * * *
V129 CONTAMINATION OF THE WATER BY V090 MAJOR CROP - COMMERCIAL GRASSES
* PAGE 1 OF 1

V129	COUNT / ROW PCT / COL PCT / TOT PCT	V090 0.	1.	2.	ROW TOTAL
NO	0.	31 70.5 77.5 40.8	9 20.5 52.9 11.8	4 9.1 21.1 5.3	44 57.9
YES	1.	9 28.1 22.5 11.8	8 25.0 47.1 10.5	15 46.9 78.9 19.7	32 42.1
COLUMN TOTAL		40 52.6	17 22.4	19 25.0	76 100.0

157

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTH NRTH JEND

* * * * * * * * * * * * * * * * * C R O S S T A B U L A T I O N O F * * * * * * * * * * * * * * * * * *
ECOSCORE BY V090 MAJOR CROP - COMMERCIAL GRASSES
* PAGE 1 OF 1

ECOSCORE	COUNT / ROW PCT / COL PCT / TOT PCT	V090 0.	1.	2.	ROW TOTAL
0.	COUNT	27	7	6	40
	ROW PCT	67.5	17.5	15.0	52.6
	COL PCT	67.5	41.2	31.6	
	TOT PCT	35.5	9.2	7.9	
1.	COUNT	13	10	13	36
	ROW PCT	36.1	27.8	36.1	47.4
	COL PCT	32.5	58.8	68.4	
	TOT PCT	17.1	13.2	17.1	
COLUMN TOTAL		40	17	19	76
		52.6	22.4	25.0	100.0

155

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * * * * * * * * * * * * * C R O S S T A B U L A T I O N O F * * * * * * * * * * * * * * * * * *
ECOSCORE BY V316 PRESENCE OF MOSQUES
* PAGE 1 OF 1

ECOSCORE	V316 COUNT / ROW PCT / COL PCT / TOT PCT	NO 0.	YES 1.	ROW TOTAL
0.	COUNT	17	23	40
	ROW PCT	42.5	57.5	52.6
	COL PCT	42.5	63.9	
	TOT PCT	22.4	30.3	
1.	COUNT	23	13	36
	ROW PCT	63.9	36.1	47.4
	COL PCT	57.5	36.1	
	TOT PCT	30.3	17.1	
COLUMN TOTAL		40	36	76
		52.6	47.4	100.0

156

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * * * * * * * * * * * * * * C R O S S T A B U L A T I O N O F * * * * * * * * * * * * * * * * * * *
ECOSCORE BY V336 # OF RETAIL SHOPS
* PAGE 1 OF 1

ECOSCORE	V336 COUNT ROW PCT COL PCT TOT PCT	1.	2.	3.	ROW TOTAL
0.		13 32.5 48.1 17.1	16 40.0 57.1 21.1	11 27.5 52.4 14.5	40 52.6
1.		14 38.9 51.9 18.4	12 33.3 42.9 15.8	10 27.8 47.6 13.2	36 47.4
COLUMN TOTAL		27 35.5	28 36.8	21 27.6	76 100.0

```
TRANSPACE REQUIRED..      600 BYTES
     6 TRANSFORMATIONS
    16 RECODE VALUES + LAG VARIABLES
    23 IF/COMPUTE OPERATIONS

CPU TIME REQUIRED..    0.78 SECONDS


            10 PEARSON CORR   ECOSCORE WITH V090,V316,V336

 ***** PEARSON CORR PROBLEM REQUIRES      144 BYTES WORKSPACE *****
```

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

- - - - - - - - - - - - - P E A R S O N C O R R E L A T I O N C O E F F I C I E N T S - - - - - - - - - - - -

	V090	V316	V336
ECOSCORE	0.3135	-0.2139	-0.0386
	(76)	(76)	(76)
	P=0.003	P=0.032	P=0.370

(COEFFICIENT / (CASES) / SIGNIFICANCE) (A VALUE OF 99.0000 IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED)

SPSS BATCH SYSTEM

01/15/82 PAGE B 1

SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM

ORDER FROM MCGRAW-HILL: SPSS, 2ND ED. (PRINCIPAL TEXT)
SPSS PRIMER (BRIEF INTRO TO SPSS)
SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)

ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS
SPSS POCKET GUIDE, RELEASE 8
KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION.. ALLOWS FOR..
WORKSPACE 17920 BYTES 25 TRANSFORMATIONS
TRANSPACE 2560 BYTES 102 RECODE VALUES + LAG VARIABLES
412 IF/COMPUTE OPERATIONS

```
1 GET FILE      KSOC
```

FILE KSOC HAS 409 VARIABLES

THE SUBFILES ARE..

NAME	NO OF CASES
WEST	24
SOUT	11
CNTR	17
NRTH	19
JEND	9

CPU TIME REQUIRED.. 0.50 SECONDS

```
2 SELECT IF     (NOT (ID EQ 102001 OR 102002 OR 103011 OR 207004))
3 COMPUTE       ECOSCORE=(V121+V122+V124+V125+V127+V129)
4 RECODE        ECOSCORE(LO THRU 4=0)(5 THRU HI=1)
5 RECODE        V090(1 THRU HI=1)
6 RECODE        V316(1 THRU HI=1)
7 CROSSTABS     TABLES=ECOSCORE BY V090 BY V316
8 STATISTICS    ALL
```

***** GIVEN WORKSPACE ALLOWS FOR 814 CELLS, 814 TABLES WITH 3 DIMENSIONS FOR CROSSTAB PROBLEM *****

150

```
FILE    RSOC      (CREATION DATE = 12/09/81)
SUBFILE   WEST      SBEI      CNTR      NRTH      JEND

* * * * * * * * * * * * * * * * * * C R O S S T A B U L A T I O N   O F  * * * * * * * * * * * * * * * * *
   ECOSCORE                                                    BY  V090      MAJOR CROP - COMMERCIAL GRASSES
CONTROLLING FOR..
   V316      PRESENCE OF MOSQUES                              VALUE =        0.  NO
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * PAGE 1 OF 1

                  V090
          COUNT  I
          ROW PCT I                   ROW
          COL PCT I                   TOTAL
          TOT PCT I      0.I      1.I
ECOSCORE  --------I--------I--------I
              0.  I    12  I     5  I    17
                  I  70.6  I  29.4  I  42.5
                  I  63.2  I  23.8  I
                  I  30.0  I  12.5  I
                 -I--------I--------I
              1.  I     7  I    16  I    23
                  I  30.4  I  69.6  I  57.5
                  I  36.8  I  76.2  I
                  I  17.5  I  40.0  I
                 -I--------I--------I
           COLUMN      19       21       40
           TOTAL     47.5     52.5    100.0

CORRECTED CHI SQUARE =     4.81229 WITH 1 DEGREE OF FREEDOM.  SIGNIFICANCE =  0.0283
      RAW CHI SQUARE =     6.31989 WITH 1 DEGREE OF FREEDOM.  SIGNIFICANCE =  0.0119
PHI =    0.39749
CONTINGENCY COEFFICIENT =    0.36938
LAMBDA (ASYMMETRIC) =  0.29412 WITH ECOSCORE DEPENDENT.        =  0.36842 WITH V090     DEPENDENT.
LAMBDA (SYMMETRIC) =  0.33333
UNCERTAINTY COEFFICIENT (ASYMMETRIC) =  0.11853 WITH ECOSCORE DEPENDENT.       =  0.11720 WITH V090     DEPENDENT.
UNCERTAINTY COEFFICIENT (SYMMETRIC) =  0.11806
KENDALL'S TAU B =     0.39749  SIGNIFICANCE =  0.0065
KENDALL'S TAU C =     0.39250  SIGNIFICANCE =  0.0065
GAMMA =    0.69163
SOMERS'S D (ASYMMETRIC) =  0.39348 WITH ECOSCORE DEPENDENT.          =  0.40153 WITH V090     DEPENDENT.
SOMERS'S D (SYMMETRIC) =  0.39747
ETA =  0.39749 WITH ECOSCORE DEPENDENT.          =  0.39749 WITH V090     DEPENDENT.
PEARSON'S R = 0.39749  SIGNIFICANCE =  0.0055
```

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

* * * * * * * * C R O S S T A B U L A T I O N O F * * * * * * * *
ECOSCORE BY V090 MAJOR CROP - COMMERCIAL GRASSES
CONTROLLING FOR..
V316 PRESENCE OF MOSQUES VALUE = 1. YES
* PAGE 1 OF 1

ECOSCORE	V090 COUNT / ROW PCT / COL PCT / TOT PCT	0.	1.	ROW TOTAL
0.	COUNT	15	8	23
	ROW PCT	65.2	34.8	63.9
	COL PCT	71.4	53.3	
	TOT PCT	41.7	22.2	
1.	COUNT	6	7	13
	ROW PCT	46.2	53.8	36.1
	COL PCT	28.6	46.7	
	TOT PCT	16.7	19.4	
COLUMN TOTAL		21	15	36
		58.3	41.7	100.0

CORRECTED CHI SQUARE = 0.58137 WITH 1 DEGREE OF FREEDOM. SIGNIFICANCE = 0.4456
RAW CHI SQUARE = 1.24185 WITH 1 DEGREE OF FREEDOM. SIGNIFICANCE = 0.2651
PHI = 0.18573
CONTINGENCY COEFFICIENT = 0.18261
LAMBDA (ASYMMETRIC) = 0.0 WITH ECOSCORE DEPENDENT. = 0.06667 WITH V090 DEPENDENT.
LAMBDA (SYMMETRIC) = 0.03571
UNCERTAINTY COEFFICIENT (ASYMMETRIC) = 0.02627 WITH ECOSCORE DEPENDENT. = 0.02529 WITH V090 DEPENDENT.
UNCERTAINTY COEFFICIENT (SYMMETRIC) = 0.02577
KENDALL'S TAU B = 0.18573 SIGNIFICANCE = 0.1359
KENDALL'S TAU C = 0.17593 SIGNIFICANCE = 0.1359
GAMMA = 0.37255
SOMERS'S D (ASYMMETRIC) = 0.18095 WITH ECOSCORE DEPENDENT. = 0.19064 WITH V090 DEPENDENT.
SOMERS'S D (SYMMETRIC) = 0.18567
ETA = 0.18573 WITH ECOSCORE DEPENDENT. = 0.18573 WITH V090 DEPENDENT.
PEARSON'S R = 0.18573 SIGNIFICANCE = 0.1391

162

```
TRANSPACE REQUIRED..      500 BYTES
     5 TRANSFORMATIONS
     7 RECODE VALUES + LAG VARIABLES
    23 IF/COMPUTE OPERATIONS

CPU TIME REQUIRED..     0.37 SECONDS

          9 FINISH          (SPSS GENERATED)

NORMAL END OF JOB.
     9 CONTROL CARDS WERE PROCESSED.
     0 ERRORS WERE DETECTED.
```

Table Relationship of eclogical problems to importance of commercial grasses, controlling on presence of mosques. (N= 76)

	Commercial grasses important*		Not important	
	Mosques absent	present	Mosques absent	present
Percent of subdistricts with 5+ ecological problems	76.2%	46.7%	36.8%	28.6%
Number subdistricts	(21)	(15)	(19)	(21)

*Combines important and very important categories

EXERCISE 8. SCATTERGRAMS AND CORRELATION

PROBLEM: How do you analyze and measure the association between two variables?

BRIEF ANSWER: Examine the relationship by way of a scattergram and then compute the correlation coefficient.

READING: SPSS Manual, pps. 276ff.

OUTPUT: A1-7; B1-2

COMMENTARY: The question and the brief answer given above assumes that you are dealing with continuous variables, such as those that you worked with in Exercise 2. If you are dealing with categorical variables, then you should make a contingency table (i.e. cross tabulation), compute percentages (which is one measure of association), and then calculate one or another coefficient of association. However, when it comes to continuous variables, the problem is both simpler and more difficult. It is simpler because there is widespread agreement that Pearson's correlation coefficient is the best measure of association for most variables. But it is more problematic, because the relationship between continuous variables can take forms that depart from linearity. It is for this reason that most analysts call for scattergrams on a range of their variables in order to be able to examine in detail the relationship.

STEPS:

1. Page A1 of the output shows the SPSS commands for getting scattergrams for three variables, the number of agglomerations, the proportion of the populations of the subdistrict that live in agglomerations, and population density. As you see, the second variable had to be computed, and you should look up the components of that computed variable.

2. Page 2 of the output shows the scattergram that slopes up to the "northeast" corner of the graph. Such a slope reflects a positive association, and indeed the correlation coefficient shown below is .75. In sociological research, that is a high correlation, inasmuch as the statistic moves from -1 through

0 to +1. but a better estimate of the strength of association is given by "R squared," which is .57. That is, you multiply the correlation coefficient by itself to get R^2, and this number can be interpreted as a 56 percent reduction in prediction error over what you could have done by chance. The third number is the level of significance, which is so high that no numbers show up in the five digits. So it is highly significant, which means that there is practically no chance that this level of association could have occurred by chance. That is all very nice, but it is practically meaningless when you are working with the universe of subdistricts, as we are here. Significance level is important if you are trying to judge whether a correlation based on a sample is likely to hold up for the whole universe. But that problem is rather special and occurs mostly in census studies or in sample survey analysis. For most purposes the analyst should pay attention to the strength of the association. Writers who exclaim about the "highly significant" correlations are simply poorly trained. Any correlation, even one as low as .05, can be statistically significant if you have enough cases.

In the second row you see the "intercept" and "slope" values. The intercept tells you where the line crosses the "y" axis, while the slope value tells you how many units on the "y" axis you have to go up for every unit of the "x" axis. Thus, in this case the slope is .108, which means that for an "x" value of 6.60 you should go up .666 units of "y", which is about a half inch up. If you determine several of those points, you can plot the "regression line," which is the best fitting line for all those dots. The intercept tells you that you should start drawing the line .398 units up the "y" axis.

3. What does it all mean? It means that if a subdistrict has a lot of agglomerations, it is likely to have a high proportion of its population

living in those agglomerations. That is almost true by definition, of course, and that is why the correlation coefficient is so high. However, notice that some districts do not conform to this rule. These are called "outliers." More concretely, they are exceptions that you must investigate. By now you should be able to figure out a way of identifying these subdistricts. It will require going back to the original listings of the values of these variables.

Page 4 is an example of the kind of messy scattergrams you are more likely to encounter in rural development research. There is still a respectable correlation, but the amount of variance explained as measured by "R^2" is only .147. However, with 76 cases it is still highly significant, whatever that means when you have the whole universe. Notice that in this case the intercept is a minus number, which means you must extend the "y" axis down in order to catch that point.

The big question is what to make of all those zero dots along the "x" axis. The subdistricts have no agglomerations, but they have varying levels of population density. Well, that is certainly possible, and it suggests that the subdistricts with no agglomerations should have been selected out of the sample before the correlation was run. The other problem is all those cases lying in the northwest corner. Although their density levels are only moderate (30-40), they have five to seven agglomerations. What we have here are some of the subdistricts with a number of family compounds distributed over the territory, some of which may be incipient subcommunities of a larger type. Subdistricts with this type of settlement pattern often have a number of springs or other localized water sources, but the rainfall is still not high.

Page 6 shows a similar scattergram, this time between the proportion of the subdistrict population living in agglomerations and the population density.

We already know that the correlation between agglomerated population and the number of agglomerations is very high, so it is somewhat surprising to find this weak correlation. The problem posed by this scattergram is the same as before: What factors other than population density could be introduced to explain these deviant cases or deviant clusters? This is the problem of explaining "statistical variance," and it is the statistical analogue of a general causal explanation. The statistical method that is appropriate to this problem is "multiple regression," because it facilitates the analysis of how a number of factors contribute to the explanation of a dependent variable. Indeed, Table 2.11 on page 51 shows a regression analysis of the number of subcommunities. In addition to population density, it introduces the location of the subdistrict (north versus other) and distance from the district headquarters. The last variable did not contribute to the explanation, but density and location explain 67 percent of the variance.

4. So the general rule in correlation analysis is to start with the examination of some scattergrams. It is the statistical equivalent of "anthropological reconnaissance." It is a method of getting to know your data in all its pecularities. But once you are familiar with it, it is convenient to present correlation coefficients in the form of matrices, as shown in the output B2. This table shows the intercorrelations of all three variables. Actually, the figures above and below the diagonal are the same, but with large matrices they facilitate finding particular correlations. In this case, it would be better to delete one-half of the matrix for use in a published table. Actually, most editors would tell you to discuss the three correlation coefficients in the text and thus save the expense of the table.

In addition to its use as a summary of interrelationships, the correlation matrix is often used as a first step in the exploration of one's data. This is known as "running everything against everything." Running a large matrix without knowing what you are looking for is bad practice if for no other reason than the fact that you will be overwhelmed with numbers. On the other hand, if you have a pretty good idea what you expect to find, then it can be an efficient technique, because it permits you to check alternative patterns quite quickly.

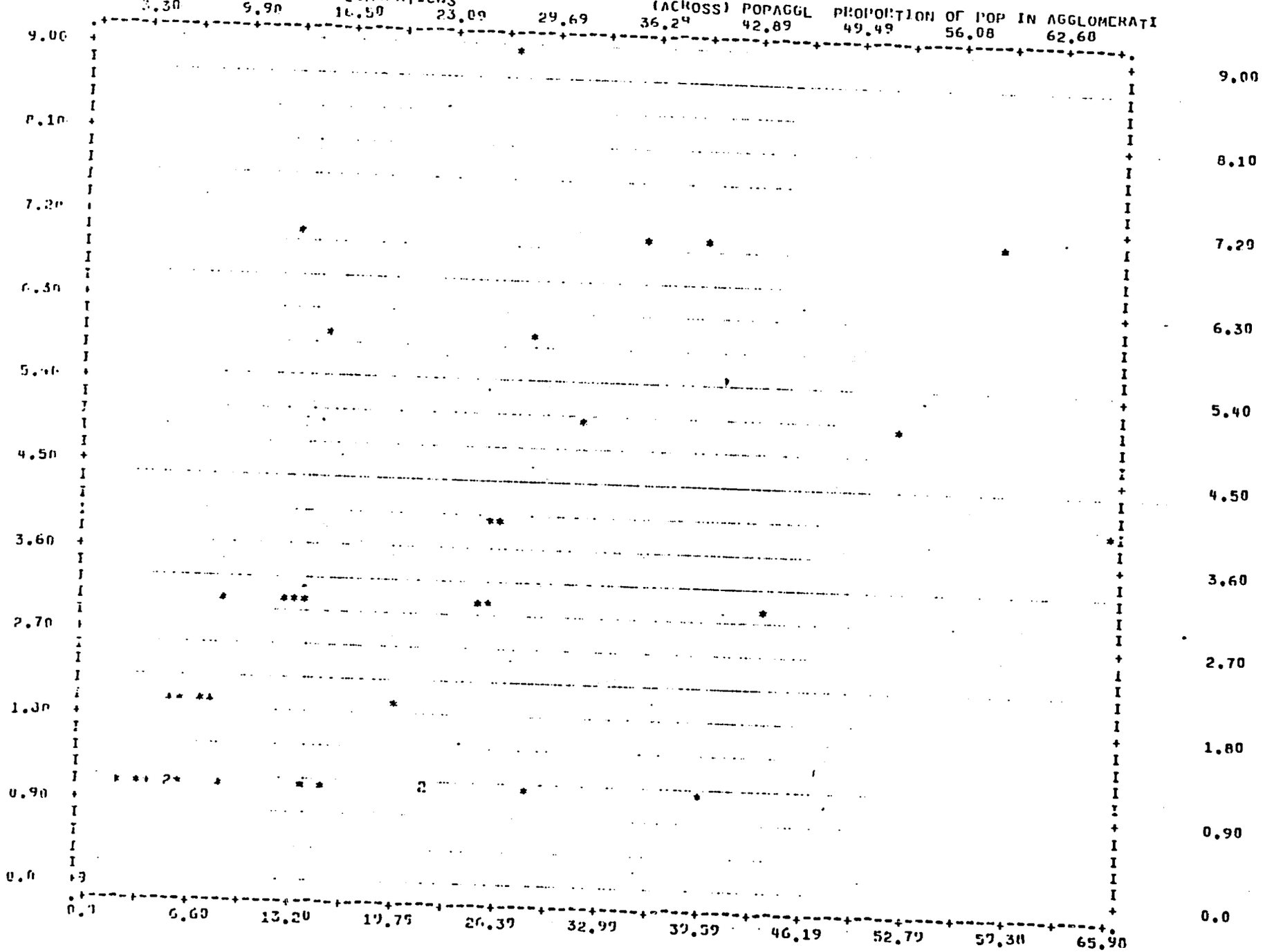
One last comment: An earlier era of sociological analysis relied quite heavily on correlation analysis (including partial correlations) as a general tool for analysis. However, this approach is not so popular these days. Analysts stay either with cross-tabular analysis (some versions of which are quite sophisticated and mathematical) or they move directly to regression. The latter method is so much more flexible in handling multiple causes, and it is easier to introduce "controls." Another reason why correlation analysis is less popular is that computers generate so many correlations that they are hard to manage. Regression is more compact.

```
*********SPSS     FT06F001 ************RN3.PR1 ***FMT <00>* *****262 LINES *** 11:19.42PM  21 DEC 81  WARREN          SUSAN BENTLEY
SPSS BATCH SYSTEM
                                                                                  12/21/81        PAGE A 1
SPSS FOR OS/360, VERSION H, RELEASE 0.1, MAY 20, 1980

                          CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM
ORDER FROM MCGRAW-HILL:  SPSS, 2ND ED. (PRINCIPAL TEXT)        ORDER FROM SPSS INC.:  SPSS STATISTICAL ALGORITHMS
                         SPSS PRIMER (BRIEF INTRO TO SPSS)                             SPSS POCKET GUIDE, RELEASE 8
                         SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 3 6)                   KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION..      ALLOWS FOR..     25 TRANSFORMATIONS
WORKSPACE      17920 BYTES                      102 RECODE VALUES + LAG VARIABLES
TRANSPACE       2560 BYTES                      412 IF/COMPUTE OPERATIONS

          1 GET FILE       RSOC


                    FILE RSOC     HAS 409 VARIABLES

                    THE SUBFILES ARE..

                              NO OF
                    NAME      CASES

                    WEST        24
                    SBLI        11
                    CNTR        17
                    NRTH        19
                    JEND         9


CPU TIME REQUIRED..     0.57 SECONDS


          2 SELECT IF      (NOT(IF EQ 102001 OR 102002 OR 103011 OR 207004))
          3 COMPUTE        POPAGGL=(V456/V449)*100
          4 ASSIGN MISSING POPAGGL(-1)
          5 VAR LABELS     POPAGGL   PROPORTION OF POP IN AGGLOMERATIONS
          6 SCATTERGRAM    V458 WITH POPAGGL/V454 WITH V452/POPAGGL WITH V452/
          7 OPTIONS        4
          8 STATISTICS     ALL

***** GIVEN WORKSPACE ALLOWS FOR   1118 CASES FOR SCATTERGRAM PROBLEM *****
```

69.

SPSS BATCH SYSTEM

FILE RSOC (CREATION DATE = 12/09/81) 12/21/81 PAGE 2
SUBFILE WEST
SCATTERGRAM OF SBEI CNTR DETH JEND
(DOWN) V454 #AGGLOMERATIONS (ACROSS) POPAGGL PROPORTION OF POP IN AGGLOMERATI



170

STATISTICS..

CORRELATION (R)-	0.75390	R SQUARED -	0.56836	SIGNIFICANCE -	0.00000
STD ERR OF EST -	1.46706	INTERCEPT (A) -	0.39824	SLOPE (B) -	0.10833
PLOTTED VALUES -	76	EXCLUDED VALUES-	0	MISSING VALUES -	0

'********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

SPSS PATCH SYSTEM

FILE USER
SUBFILE VEST
SCATTERGRAM OF

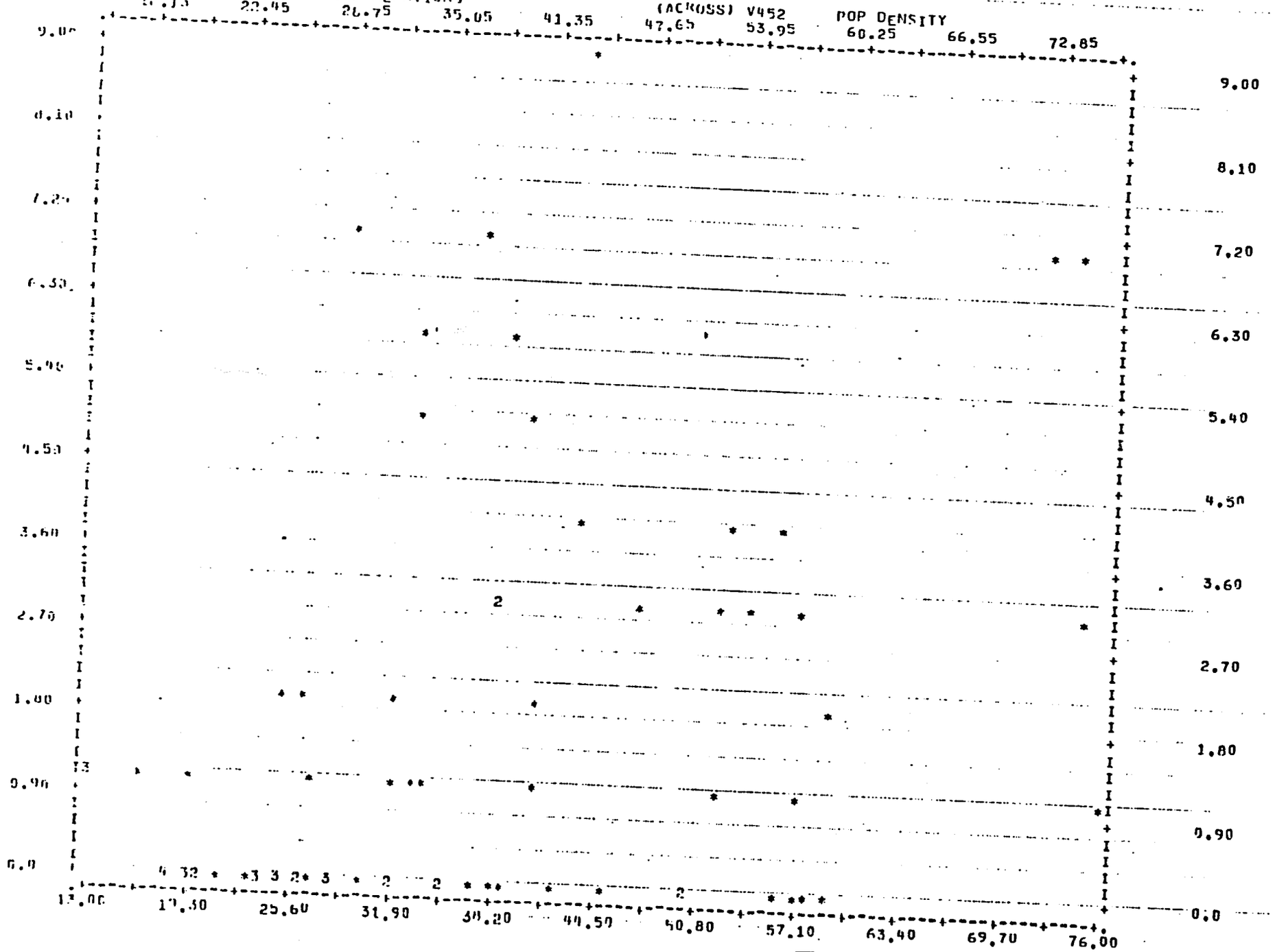
CREATION DATE = 12/09/81)

12/21/01 PAGE 4

SPEI (DOWN) V450 CNTR NRTH JEND

#AGGLOMERATIONS

(ACROSS) V452 POP DENSITY

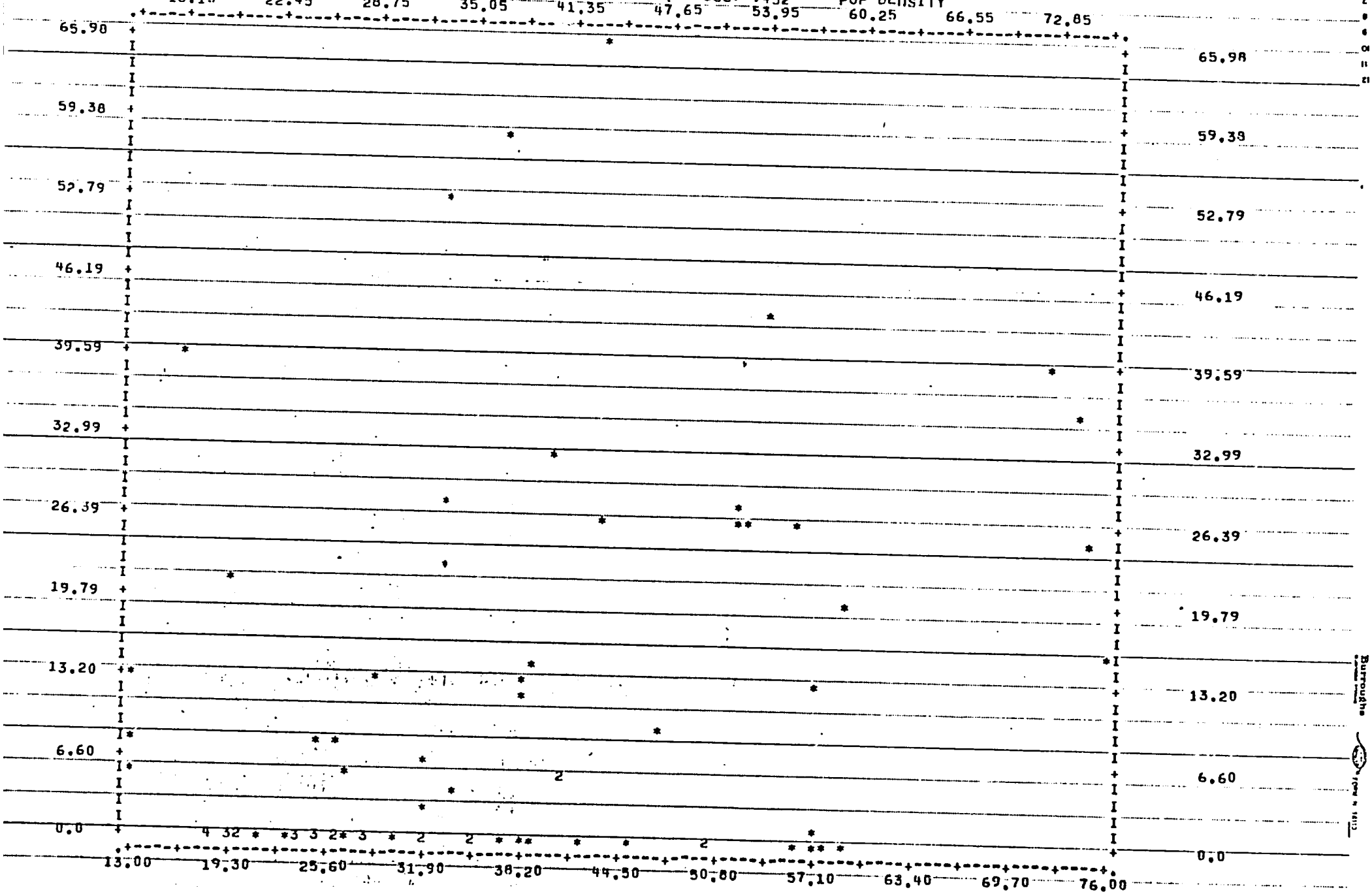


STATISTICS..

CORRELATION (R)-	0.38307	R SQUARED -	0.14736	SIGNIFICANCE -	0.00031
STD ERR OF EST -	2.06192	INTERCEPT (A) -	-0.43749	SLOPE (B) -	0.05411
PLOTTED VALUES -	76	EXCLUDED VALUES-	0	MISSING VALUES -	0

'********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND
SCATTERGRAM OF (DOWN) POPAGGL PROPORTION OF POP IN AGGLOMERATI (ACROSS) V452 POP DENSITY



STATISTICS..

CORRELATION (R) -	0.34978	R SQUARED	-	0.12235	SIGNIFICANCE	-	0.00098
STD ERR OF EST -	14.55871	INTERCEPT (A)	-	-2.07779	SLOPE (B)	-	0.34311
PLOTTED VALUES -	76	EXCLUDED VALUES-		0	MISSING VALUES	-	0

'********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM

ORDER FROM MCGRAW-HILL: SPSS, 2ND ED. (PRINCIPAL TEXT)
SPSS PRIMER (BRIEF INTRO TO SPSS)
SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)

ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS
SPSS POCKET GUIDE, RELEASE 8
KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION.. ALLOWS FOR..
WORKSPACE 17920 BYTES — 25 TRANSFORMATIONS
TRANSPACE 2560 BYTES — 102 RECODE VALUES + LAG VARIABLES
412 IF/COMPUTE OPERATIONS

```
  1 SET FILE      RSOC
```

FILE RSOC HAS 409 VARIABLES

THE SUBFILES ARE..

NAME	NO OF CASES
WEST	24
SBEI	11
CNTR	17
NRTH	19
JFND	9

CPU TIME REQUIRED.. 0.50 SECONDS

```
  2 SELECT IF      (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
  3 COMPUTE        POPAGGL=(V456/V449)*100
  4 ASSIGN MISSING POPAGGL(-1)
  5 VAR LABELS     POPAGGL   PROPORTION OF POP IN AGGLOMERATIONS
  6 PEARSON CORR   V452,V454,POPAGGL
```

***** PEARSON CORR PROBLEM REQUIRES 144 BYTES WORKSPACE *****

- - - - - - - - - - - - - P E A R S O N C O R R E L A T I O N C O E F F I C I E N T S - - - - - - - - - - - -

	V452	V454	POPAGGL
V452	1.0000 (0) P=*****	0.3639 (76) P=0.000	0.3498 (76) P=0.001
V454	0.3639 (76) P=0.000	1.0000 (0) P=*****	0.7539 (76) P=0.000
POPAGGL	0.3498 (76) P=0.001	0.7539 (76) P=0.000	1.0000 (0) P=*****

(COEFFICIENT / (CASES) / SIGNIFICANCE) (A VALUE OF 99.0000 IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED)

EXERCISE 9. TRANSFORMATIONS AND SPECIAL COEFFICIENTS

Problem: What do you do with a variable if it is so skewed that your statistical analysis is likely to be thrown off?

Brief Answer: Change the distribution of the variable by "transforming" it.

Reading: A good treatment may be found in Rummel (1970).

Output: A1-5; B1-7; C1.

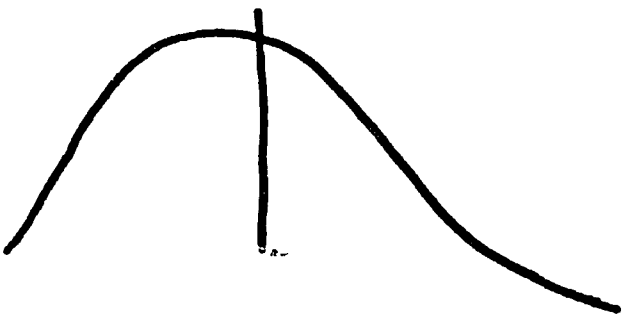
Commentary: There are two reasons for transforming: First, transformations are sometimes necessary so that a distribution will fit the assumptions of the statistics that you are using. Most statistics assume a "normal" distribution, so if your distribution departs very much from normality, the results will be weaker and perhaps misleading. The second reason for transforming distributions is to eliminate random or unique values which may throw off your analysis. The problem of dealing with "outliers" falls in this category.

On the following page is a diagram of two general types of skewness. The rule of thumb is that if the coefficient of skewness is greater than -2 or +2, it is wise to transform. However, some analysts use 3 as their warning signal. There are two general types of transformations: (a) multiplying by a number such as a logarithm or the number itself (squaring) or (b) "grouping." Grouping is simply collapsing down extreme values. For example, if the highest value of a distribution is 350 but the next lowest is 100, then it is possible to recode the 350 to read 100. In this case, you collapse down to the next highest number.

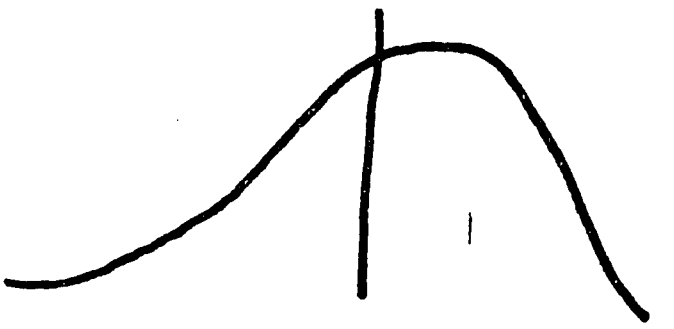
In using type 'a' transformations, a standard procedure is to correct right-hand skews by multiplying by the log of 10 and left-hand skews by squaring. There are a lot of other possibilities depending on how extreme the skew is, but these usually work.

The instructions for regrouping will be picked up later.

Right skew
(Positive sign)



Left skew
(Negative sign)



RECTIFICATION:

Type a	logarithm to base 10 (log X_j)	squaring
Type b	grouping	grouping

Having explained the rationale for transformations, the next thing to say is to avoid them if at all possible. They can improve the statistics, it is true, but they certainly complicate the interpretation. It is not easy to explain to people what a correlation of .60 between two transformed variables really means. Some analysts do the transformations and run their correlations to make sure that the result is correct. Then, if the raw variables give them adequate correlations, they use those for actually presenting the data.

Steps:

1. The exercise shows how to transform three variables. One is squared and the other two are multiplied by log 10. Line three shows how to square, and lines four and five show the log transformations.

On page 4, you see the condescriptive for the number of grades in primary schools. Skewness is -3.994. Down below is the transformed value, called TGRADES, which reduces the skewness to -2.186. Notice that kurtosis is also reduced, and that is a good thing, but the transformations are not designed to deal with this problem. Kurtosis, by the way, is the "peakedness" of the distribution. Imagine a cowboy's hat. The taller the hat, the greater the kurtosis, and the wider the brim to one side or the other, the greater the skewness.

V022, the number of merchants in the subdistrict, is skewed to the right and must be corrected by a log transformation. The result is shown below where the skewness is reduced to .918. Similarly, the number of retail stores is reduced to .31. Needless to say, the means of the distribution also change because you are now dealing with the logarithm of the original values.

2. What about grouping? The list of cases on page 2 will show you whether there any extreme values, and indeed there are. Case 57 has a value of 350 for variable V022, number of merchants. This is probably some kind of error. It looks as if an extra 0 got added somehow. The next highest number

```
*********SPSS     FT06F001 ************RM3.PR1 ***FMT <00>* *****207 LINES *** 12:56.16AM  12 JAN 82  WARREN          SUSAN BENTLEY
SPSS BATCH SYSTEM                                                                    01/12/82        PAGE A 1
SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

                                CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM
ORDER FROM MCGRAW-HILL:  SPSS, 2ND ED. (PRINCIPAL TEXT)        ORDER FROM SPSS INC.:  SPSS STATISTICAL ALGORITHMS
                         SPSS PRIMER (BRIEF INTRO TO SPSS)                             SPSS POCKET GUIDE, RELEASE 8
                         SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)                   KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION..     ALLOWS FOR..    25 TRANSFORMATIONS
WORKSPACE    17920 BYTES                      102 RECODE VALUES + LAG VARIABLES
TRANSPACE     2560 BYTES                      412 IF/COMPUTE OPERATIONS

          1 GET FILE     RSOC


                    FILE RSOC     HAS 409 VARIABLES

                    THE SUBFILES ARE..

                              NO OF
                    NAME      CASES

                    WEST        24
                    SBEI        11
                    CNTR        17
                    NRTH        19
                    JEND         9

CPU TIME REQUIRED..    0.46 SECONDS


          2 SELECT IF       (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
          3 COMPUTE         TGRADES=V004**2
          4 COMPUTE         TMERCHANT=LG10(V022)
          5 COMPUTE         TSTORES=LG10(V336)
          6 ASSIGN MISSING  TGRADES TO TSTORES(-1)
          7 LIST CASES      CASES=80/VARIABLES=V004,V022,V336
          8 CONDESCRIPTIVE  V004,V022,V336,TGRADES,TMERCHANT,TSTORES

***** GIVEN WORKSPACE ALLOWS FOR  358 VARIABLES FOR CONDESCRIPTIVE PROBLEM *****

          9 STATISTICS      ALL
```

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

CASE-N	V004	V022	V336
1	6.	25.	0.
2	6.	15.	0.
3	6.	1.	3.
4	6.	16.	50.
5	6.	2.	0.
6	6.	0.	0.
7	6.	6.	0.
8	6.	8.	2.
9	6.	60.	0.
10	6.	7.	10.
11	6.	17.	0.
12	6.	6.	4.
13	6.	4.	6.
14	6.	10.	4.
15	6.	4.	2.
16	6.	10.	6.
17	8.	5.	5.
18	6.	8.	1.
19	6.	999.	0.
20	0.	3.	0.
21	6.	0.	9.
22	6.	30.	6.
23	6.	2.	2.
24	6.	15.	2.
25	6.	3.	3.
26	6.	7.	7.
27	6.	4.	4.
28	6.	12.	5.
29	6.	2.	2.
30	9.	6.	0.
31	6.	6.	6.
32	6.	2.	2.
33	0.	0.	3.
34	6.	3.	3.
35	6.	6.	6.
36	6.	1.	1.
37	6.	5.	5.
38	3.	18.	85.
39	6.	5.	5.
40	6.	10.	10.
41	6.	6.	5.
42	6.	5.	5.
43	6.	8.	14.
44	6.	3.	1.
45	6.	5.	5.

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

CASE-N	V004	V022	V336
46	6.	30.	9.
47	6.	0.	7.
48	6.	0.	6.
49	6.	5.	5.
50	6.	8.	8.
51	6.	10.	10.
52	6.	1.	1.
53	6.	20.	20.
54	6.	12.	12.
55	6.	12.	12.
56	6.	10.	4.
57	6.	350.	12.
58	6.	5.	5.
59	6.	3.	4.
60	4.	4.	4.
61	6.	11.	5.
62	6.	4.	8.
63	6.	0.	2.
64	6.	10.	8.
65	6.	13.	13.
66	6.	2.	2.
67	7.	0.	10.
68	6.	6.	6.
69	3.	8.	7.
70	6.	4.	4.
71	6.	4.	4.
72	6.	0.	1.
73	6.	70.	12.
74	6.	15.	15.
75	6.	2.	13.
76	6.	10.	10.

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBEI CNTR NRTH JEND

VARIABLE V004 NUMBER OF GRADES IN PRIMARY SCHOOL

MEAN	5.773	STD ERROR	0.131	STD DEV	1.134
VARIANCE	1.286	KURTOSIS	17.948	SKEWNESS	-3.994
RANGE	8.000	MINIMUM	0.0	MAXIMUM	8.000
SUM	433.000				

VALID OBSERVATIONS - 75 MISSING OBSERVATIONS - 1

- -

VARIABLE V022 # OF MERCHANTS

MEAN	13.333	STD ERROR	4.736	STD DEV	41.015
VARIANCE	1682.252	KURTOSIS	63.485	SKEWNESS	7.727
RANGE	350.000	MINIMUM	0.0	MAXIMUM	350.000
SUM	1000.000				

VALID OBSERVATIONS - 75 MISSING OBSERVATIONS - 1

- -

VARIABLE V336 # OF RETAIL SHOPS

MEAN	6.947	STD ERROR	1.291	STD DEV	11.252
VARIANCE	126.611	KURTOSIS	33.484	SKEWNESS	5.321
RANGE	85.000	MINIMUM	0.0	MAXIMUM	85.000
SUM	528.000				

VALID OBSERVATIONS - 76 MISSING OBSERVATIONS - 0

- -

VARIABLE TGRADES

MEAN	34.600	STD ERROR	0.973	STD DEV	8.428
VARIANCE	71.027	KURTOSIS	10.340	SKEWNESS	-2.186
RANGE	64.000	MINIMUM	0.0	MAXIMUM	64.000
SUM	2595.000				

VALID OBSERVATIONS - 75 MISSING OBSERVATIONS - 1

- -

FILE RSOC (CREATION DATE = 12/09/81)
SUBFILE WEST SBE1 CNTR NRTH JEND

VARIABLE TMERCHAN

MEAN	0.842	STD ERROR	0.053	STD DEV	0.435
VARIANCE	0.189	KURTOSIS	2.838	SKEWNESS	0.918
RANGE	2.544	MINIMUM	0.0	MAXIMUM	2.544
SUM	56.383				

VALID OBSERVATIONS - 67 MISSING OBSERVATIONS - 9

- -

VARIABLE TSTORES

MEAN	0.720	STD ERROR	0.046	STD DEV	0.372
VARIANCE	0.138	KURTOSIS	1.270	SKEWNESS	0.314
RANGE	1.929	MINIMUM	0.0	MAXIMUM	1.929
SUM	47.519				

VALID OBSERVATIONS - 66 MISSING OBSERVATIONS - 10

SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

ORDER FROM MCGRAW-HILL: CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM
SPSS, 2ND ED. (PRINCIPAL TEXT)
SPSS PRIMER (BRIEF INTRO TO SPSS)
SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)

ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS
SPSS POCKET GUIDE, RELEASE 8
KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION.. ALLOWS FOR..
WORKSPACE 179,0 BYTES 25 TRANSFORMATIONS
TRANSPACE 25,0 BYTES 102 RECODE VALUES + LAG VARIABLES
412 IF/COMPUTE OPERATIONS

```
1 GET FILE      RSOC
```

FILE RSOC HAS 409 VARIABLES

THE SUBFILES ARE..

NAME	NO OF CASES
WEST	24
SBEI	11
CNTR	17
NRTH	15
JEND	5

CPU TIME REQUIRED.. 0.58 SECONDS

```
2 SELECT IF      (NOT(ID EQ 102001 OR 102002 OR 103011 OR 207004))
3 RUN SUBFILES   (WEST) (CNTR) (NRTH) (SBEI,JEND)
4 CONDESCRIPTIVE V452
```

***** GIVEN WORKSPACE ALLOWS FOR 358 VARIABLES FOR CONDESCRIPTIVE PROBLEM *****

```
5 STATISTICS     1,5
```

SPSS BATCH SYSTEM

FILE RSOC (CREATION DATE = 12/09/81) 01/13/82 PAGE 2
SUBFILE WEST

VARIABLE V452 POP DENSITY

MEAN 31.045 STD DEV 13.671

VALID OBSERVATIONS - 22 MISSING OBSERVATIONS - 0

SPSS BATCH SYSTEM

FILE RSOC (CREATION DATE = 12/09/81) 01/13/82 PAGE 3
SUBFILE CNTR

VARIABLE V452 POP DENSITY

MEAN 39.118 STD DEV 16.205

VALID OBSERVATIONS - 17 MISSING OBSERVATIONS - 0

SPSS BATCH SYSTEM

FILE RSOC (CREATION DATE = 12/09/81) 01/13/82 PAGE 4
SUBFILE MRTH

VARIABLE V452 POP DENSITY

MEAN 40.611 STD DEV 16.468

VALID OBSERVATIONS - 18 MISSING OBSERVATIONS - 0

SPSS BATCH SYSTEM

FILE RSOC (CREATION DATE = 12/09/81) 01/13/82 PAGE 5
SUBFILE SEEI JEND

VARIABLE V452 POP DENSITY

MEAN 34.789 STD DEV 16.195

VALID OBSERVATIONS - 19 MISSING OBSERVATIONS - 0

```
TRANSPACE REQUIRED..      100 BYTES
      1 TRANSFORMATIONS
      0 RECODE VALUES + LAG VARIABLES
     16 IF/COMPUTE OPERATIONS

CPU TIME REQUIRED..    0.49 SECONDS


                6 RUN SUBFILES    ALL
                7 CONDESCRIPTIVE V452

***** GIVEN WORKSPACE ALLOWS FOR  358 VARIABLES FOR CONDESCRIPTIVE PROBLEM *****

                8 STATISTICS      1,5
```



```
FILE    RSOC      (CREATION DATE = 12/09/81)
SUBFILE   WEST        SOEI        CNTR        NRTH        JEND

VARIABLE  V452        POP DENSITY

MEAN            36.053                STD DEV          15.737

VALID OBSERVATIONS -       76                MISSING OBSERVATIONS -       0
```

```
*********SPSS      FT06F001 ************RM3.PR1 ***FMT <00>* ******52 LINES ***  1:40.08PM  13 JAN 82  WARREN          BENTLEY SUSAN
SPSS BATCH SYSTEM                                                                                  01/13/82      PAGE  1
SPSS FOR OS/360, VERSION H, RELEASE 8.1, MAY 20, 1980

ORDER FROM MCGRAW-HILL:  CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM
                         SPSS, 2ND ED. (PRINCIPAL TEXT)          ORDER FROM SPSS INC.:  SPSS STATISTICAL ALGORITHMS
                         SPSS PRIMER (BRIEF INTRO TO SPSS)                              SPSS POCKET GUIDE, RELEASE 8
                         SPSS UPDATE (USE W/SPSS,2ND FOR REL. 7 & 8)                    KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION..      ALLOWS FOR..     25 TRANSFORMATIONS
WORKSPACE      17920 BYTES                     102 RECODE VALUES + LAG VARIABLES
TRANSPACE       2560 BYTES                     412 IF/COMPUTE OPERATIONS

          1 GET FILE        KS6C

                         FILE KS0C      HAS 409 VARIABLES

                         THE SUBFILES ARE..

                                   NO OF
                         NAME      CASES

                         WEST        24
                         SBEI        11
                         CNTR        17
                         NRTH        19
                         JEND         5

CPU TIME REQUIRED..     0.61 SECONDS

          2 COMPUTE        CVARW=13.671/31.045
          3 COMPUTE        CVARC=16.205/39.118
          4 COMPUTE        CVARN=16.468/40.611
          5 COMPUTE        CVARSE=16.195/34.789
          6 COMPUTE        CVARALL=15.737/36.053
          7 WRITE CASES    (5(F5.4,4X))CVARW TO CVARALL

*********SPSS      FT09F001 ************RM3.PR1 ***FMT

 4404      4143      4055      4655      4365
 4404      4143      4055      4655      4365
 4404      4143      4055      4655      4365
 4404      4143      4055      4655      4365
 4404      4143      4055      4655      4365
 4404      4143      4055      4655      4365
 4404      4143      4055      4655      4365
 4404      4143      4055      4655      4365
 4404      4143      4055      4655      4365
```

139