Overview of Exercise 1

PART A: Creating a morbseq variable to explore inpatient data

Step 1. Open vashmds file.

Step 2. Determine record number in data set.

Step 3. Create yearsep variable and determine number of separations by year.

Step 4. Create a morbseq variable.

Step 5. Use morbseq to determine number of patients in dataset and number of patients with single or multiple hospital records.

Step 6. Determine mean patient age at first admission.

Step 7. Use aggregate command (or SAS and Stata equivalent) to determine the mean, median and maximum number of hospital records per patient.

PART B: Converting a type II to a type I file

Step 8. Open vascancer file.

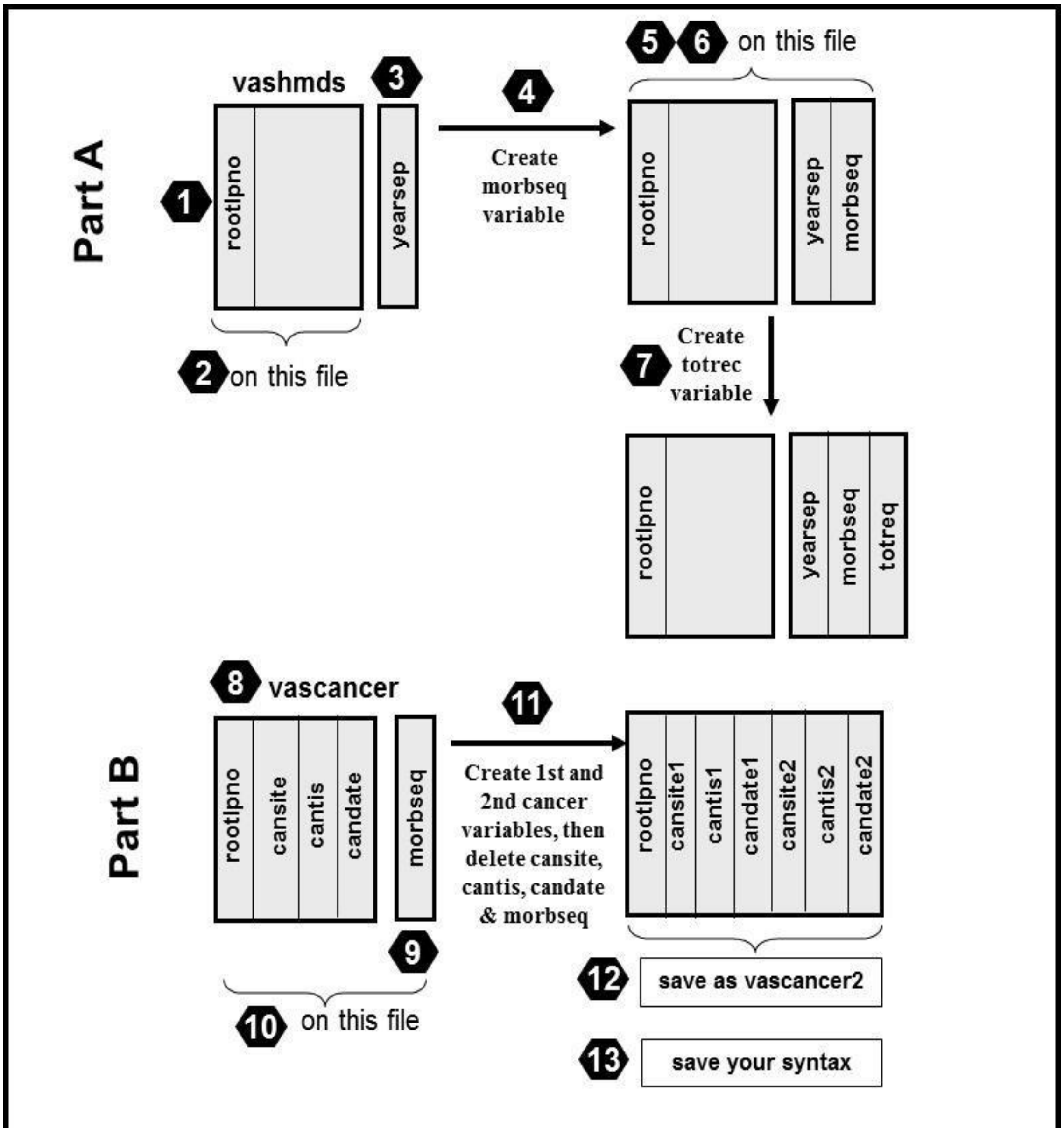
Step 9. Create a morbseq variable.

Step 10. Assess distribution of morbseq.

Step 11. Reconstruct the file as one record per individual.

Step 12. Save the reconstructed file as vascancer2.

Step 13. Save your syntax.



Exercise 1 flow diagram

Exercise Instructions

IMPORTANT: Please note that the methods you will use in this exercise are not necessarily the approach that you will take once you are more experienced and have more skills with writing syntax.

PART A: Creating a morbseq variable to explore inpatient data

In Part A of this exercise you will use syntax to explore the hospital morbidity data set. You will also calculate some basic descriptive parameters for the study sample from this file. Your aim is to create syntax to calculate basic summary statistics to complete the following table.

Characteristic	Result
Number of records in data set	
Year range of records in data set	
Number of people in data set	
Number of people with only a single hospital record	
Number of people with multiple hospital records	
Mean (±SD) age at first-time admission	
Age range at first-time admission	
Mean number of admissions per patient	
Median number of admissions per patient	
Maximum number of admissions per patient	

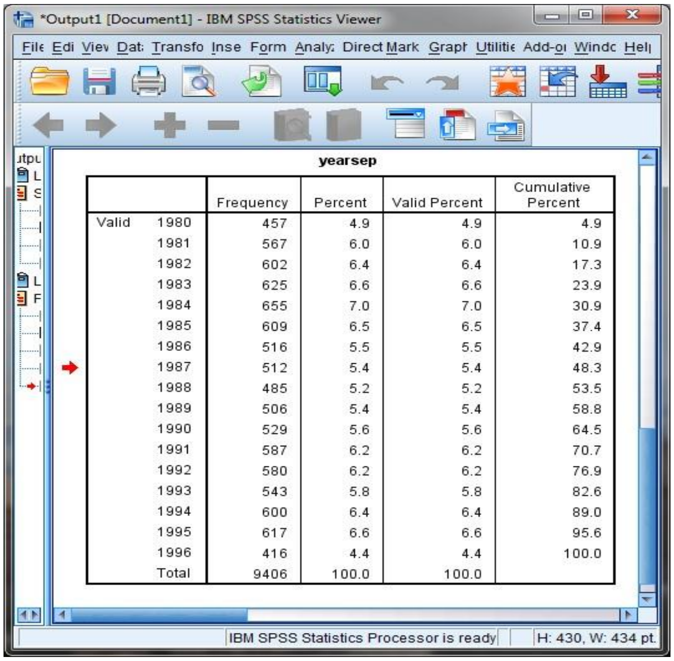
1. Open the vashmds data file in your preferred statistical software package.
2. Compose and execute syntax that will return the total number records in the vashmds file. Check that you have 9406 records. Note that for small files this can be achieved easily by scrolling to (or searching for) the end record in the data set. However, when working with large data sets containing thousands or millions of records this is not always so easily achieved.

Useful tip: The use of syntax to return total case number in a data set

SPSS

is a quick way to check you are working on file with the correct number of records, especially after you have cut files down in size with previous analyses.

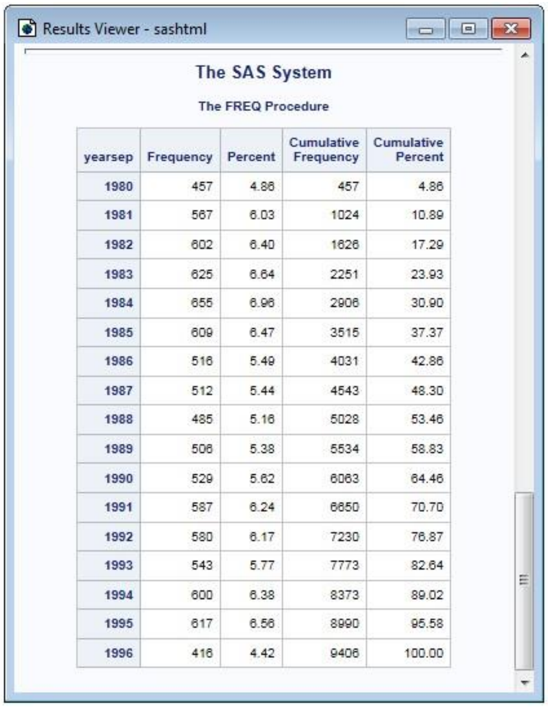
3. Now compose and execute syntax that will create a new variable called yearsep, derived from sepdate using a date function to extract the year. Run a frequency on yearsep to look at the number of hospital separations per year across the study period. The results should look like the output in the adjacent tables. Do the results look like what you would expect? What decisions might be necessary before proceeding with any further analysis?


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
IBM SPSS Statistics Processor is ready
H: 430, W: 434 pt.

yearsep

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1980	457	4.9	4.9	4.9
	1981	567	6.0	6.0	10.9
	1982	602	6.4	6.4	17.3
	1983	625	6.6	6.6	23.9
	1984	655	7.0	7.0	30.9
	1985	609	6.5	6.5	37.4
	1986	516	5.5	5.5	42.9
	1987	512	5.4	5.4	48.3
	1988	485	5.2	5.2	53.5
	1989	506	5.4	5.4	58.8
	1990	529	5.6	5.6	64.5
	1991	587	6.2	6.2	70.7
	1992	580	6.2	6.2	76.9
	1993	543	5.8	5.8	82.6
	1994	600	6.4	6.4	89.0
	1995	617	6.6	6.6	95.6
	1996	416	4.4	4.4	100.0
	Total	9406	100.0	100.0	

SAS

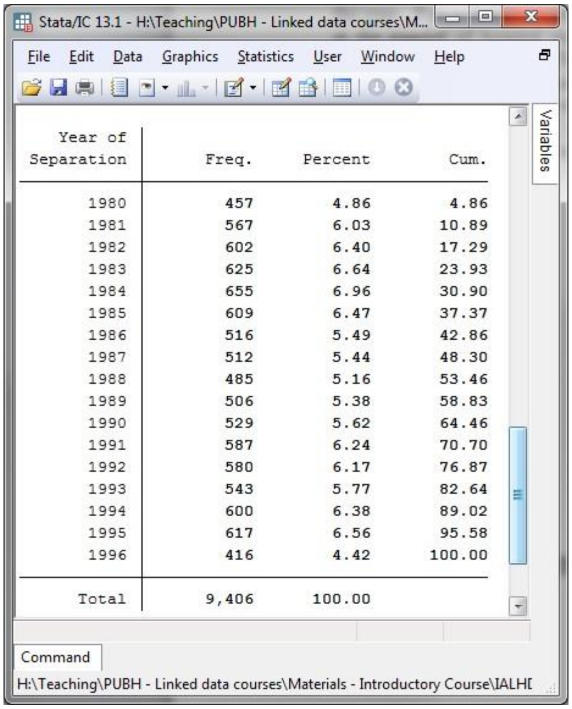


The SAS System

The FREQ Procedure

yearsep	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1980	457	4.86	457	4.86
1981	567	6.03	1024	10.89
1982	602	6.40	1626	17.29
1983	625	6.64	2251	23.93
1984	655	6.96	2906	30.90
1985	609	6.47	3515	37.37
1986	516	5.49	4031	42.86
1987	512	5.44	4543	48.30
1988	485	5.16	5028	53.46
1989	506	5.38	5534	58.83
1990	529	5.62	6063	64.46
1991	587	6.24	6650	70.70
1992	580	6.17	7230	76.87
1993	543	5.77	7773	82.64
1994	600	6.38	8373	89.02
1995	617	6.56	8990	95.58
1996	416	4.42	9406	100.00

Stata


Stata/IC 13.1 - H:\Teaching\PUBH - Linked data courses\M...
File
Edit
Data
Graphics
Statistics
User
Window
Help
Variables
Command
H:\Teaching\PUBH - Linked data courses\Materials - Introductory Course\IALHI

```
    Year of
 Separation |      Freq.     Percent        Cum.
------------+-----------------------------------
       1980 |        457        4.86        4.86
       1981 |        567        6.03       10.89
       1982 |        602        6.40       17.29
       1983 |        625        6.64       23.93
       1984 |        655        6.96       30.90
       1985 |        609        6.47       37.37
       1986 |        516        5.49       42.86
       1987 |        512        5.44       48.30
       1988 |        485        5.16       53.46
       1989 |        506        5.38       58.83
       1990 |        529        5.62       64.46
       1991 |        587        6.24       70.70
       1992 |        580        6.17       76.87
       1993 |        543        5.77       82.64
       1994 |        600        6.38       89.02
       1995 |        617        6.56       95.58
       1996 |        416        4.42      100.00
------------+-----------------------------------
      Total |      9,406      100.00
```

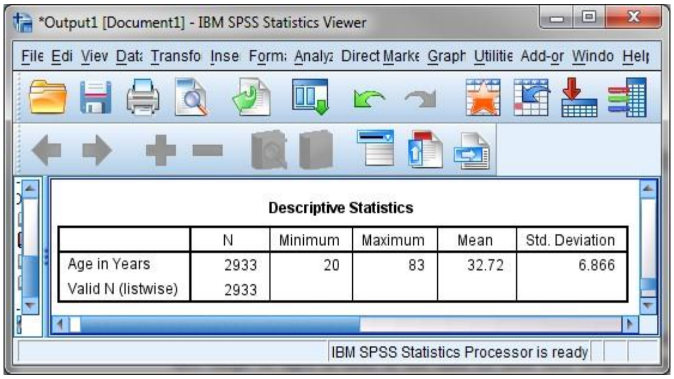
4. Now that you have explored the number of records in the data set, you should investigate some basic descriptive characteristics for the number of patients. Compose and execute syntax that will create a new variable called morbseq (for *morbidity sequence*) in the EOR loading area, which is assigned the value ‘1’ for the first hospital record for an individual, ‘2’ for the second record in the same individual, and so on. **SAS users**: you may employ by-group processing. Visually inspect the file to ensure that your syntax appears to have produced the expected result.

   Useful tip: always visually inspect the data after executing syntax to check that the observed and expected results are the same. For complex syntax, it is wise to check the results for several different individuals with variations in characteristics that required the syntax to operate in different ways.

5. Using your newly created morbseq variable, select cases where morbseq=1 and run a frequency to output the number of patients in the dataset. Check that you have 2933 patients with at least one record in the data set. Now use the morbseq variable to determine the number of patients with multiple hospital separations in the data set. There should be 1772 with multiple hospital separation records during the observation period, leaving 1161 patients with a single record.

6. Now reselect cases where morbseq=1 and run a simple descriptive analysis to obtain the mean, standard deviation and range of ages in men at the time of their first hospital separation in the data set. Your results should be as shown in the output tables below.

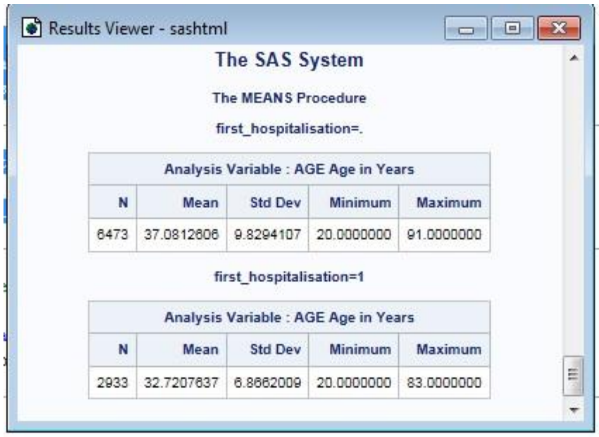
SPSS


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edi View Dat Transfo Inse Form Analy Direct Marke Graph Utilitie Add-or Windo Help
IBM SPSS Statistics Processor is ready

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age in Years	2933	20	83	32.72	6.866
Valid N (listwise)	2933				

SAS



The SAS System

The MEANS Procedure

first_hospitalisation=.

Analysis Variable : AGE Age in Years

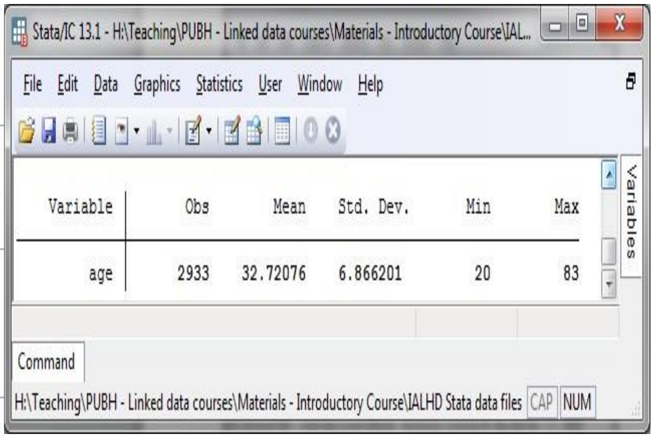
N	Mean	Std Dev	Minimum	Maximum
6473	37.0812606	9.8294107	20.0000000	91.0000000

first_hospitalisation=1

Analysis Variable : AGE Age in Years

N	Mean	Std Dev	Minimum	Maximum
2933	32.7207637	6.8662009	20.0000000	83.0000000

Stata

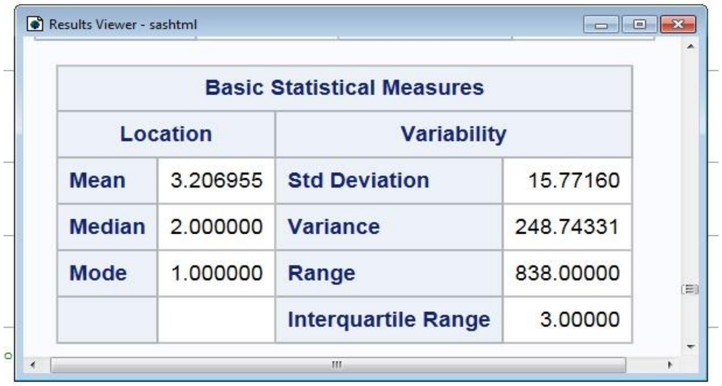

Stata/IC 13.1 - H:\Teaching\PUBH - Linked data courses\Materials - Introductory Course\IAL...
File Edit Data Graphics Statistics User Window Help
Variables
Command
H:\Teaching\PUBH - Linked data courses\Materials - Introductory Course\IALHD Stata data files
CAP
NUM

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |      2933    32.72076    6.866201         20         83
```

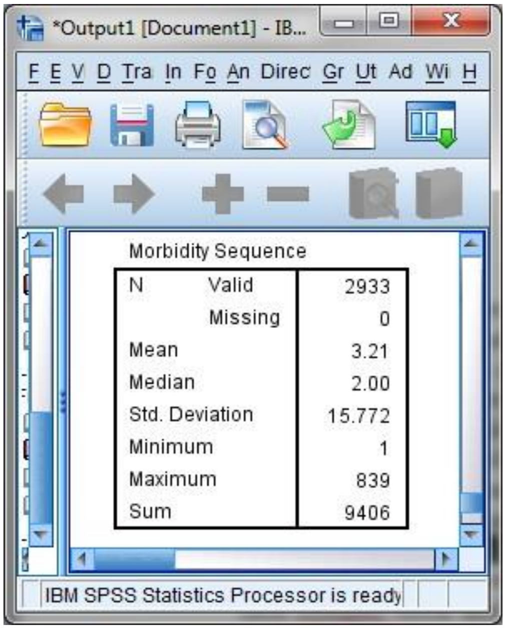
7. Now compose and execute syntax to derive the mean, median and maximum number of admissions per patient in the data set. There are multiple ways you can do this which will be explored over the course, however you will implement one specific simple approach for this exercise. **SPSS users**: the simplest approach is to use the aggregate command. **Stata users**: use the egen command. **SAS users**: you will need to create syntax using the retain statement.

Use the relevant software package command(s) to generate a new variable called totrec that returns the highest morbseq number for each patient to each record of that particular patient. Now select where morbseq=1 and calculate the mean, median and maximum number of separations per patient. If you have done this correctly, your results should look like the output tables below.

SAS



Basic Statistical Measures			
Location		**Variability**	
Mean	3.206955	**Std Deviation**	15.77160
Median	2.000000	**Variance**	248.74331
Mode	1.000000	**Range**	838.00000
		Interquartile Range	3.00000



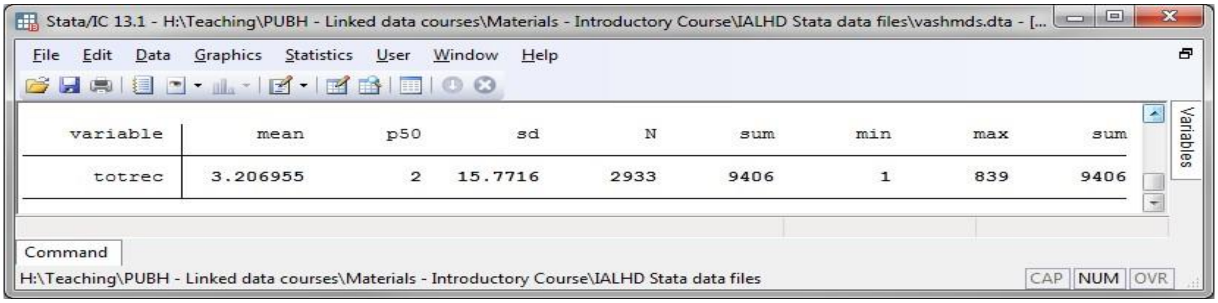
*Output1 [Document1] - IB...
F E V D Tra In Fo An Direc Gr Ut Ad Wi H
IBM SPSS Statistics Processor is ready

Morbidity Sequence

N	Valid	2933
	Missing	0
Mean		3.21
Median		2.00
Std. Deviation		15.772
Minimum		1
Maximum		839
Sum		9406

SPSS

Stata


Stata/IC 13.1 - H:\Teaching\PUBH - Linked data courses\Materials - Introductory Course\IALHD Stata data files\vashmds.dta - [...
File
Edit
Data
Graphics
Statistics
User
Window
Help
Variables
Command
H:\Teaching\PUBH - Linked data courses\Materials - Introductory Course\IALHD Stata data files
CAP
NUM
OVR

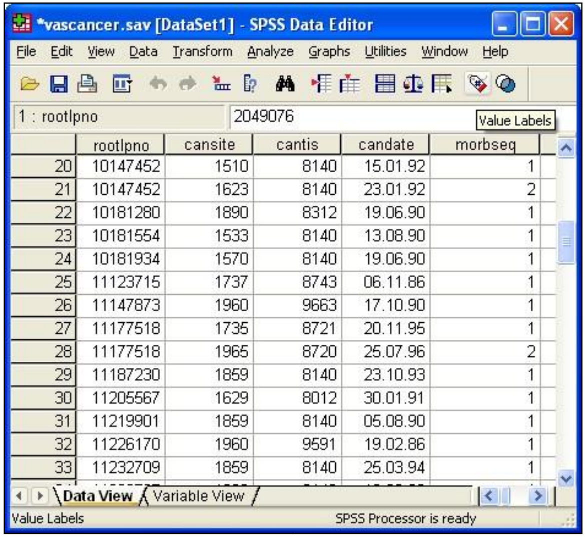
```
    variable |      mean       p50        sd         N       sum       min       max       sum
-------------+------------------------------------------------------------------------------------------
      totrec |  3.206955         2   15.7716      2933      9406         1       839      9406
--------------------------------------------------------------------------------------------------------
```

PART B: Converting a type II to a type I file

8. Open the correct vascancer data file for your preferred statistical software package.

9. Compose and execute syntax that will create a new variable called morbseq (for *morbidity sequence*) in the EOR loading area, which is assigned the value '1' for the first cancer for an individual, '2' for the second cancer in the same individual, and so on. **SAS users** : you may employ by-group processing. Visually inspect the file (especially rootlpno 11177518) to ensure that your syntax appears to have produced the expected result. **Useful tip** : always visually inspect the data after executing syntax to check that the observed and expected results are the same. For complex syntax, it is wise to check the results for several different individuals with variations in characteristics that required the syntax to operate in different ways.

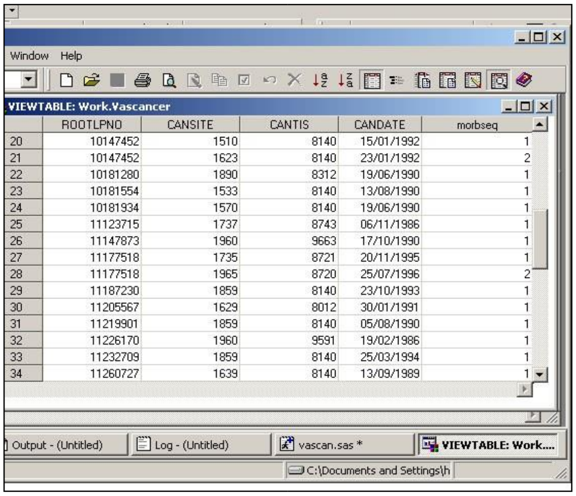
SPSS


*vascancer.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Value Labels
Data View
Variable View
Value Labels
SPSS Processor is ready

1 : rootlpno 2049076

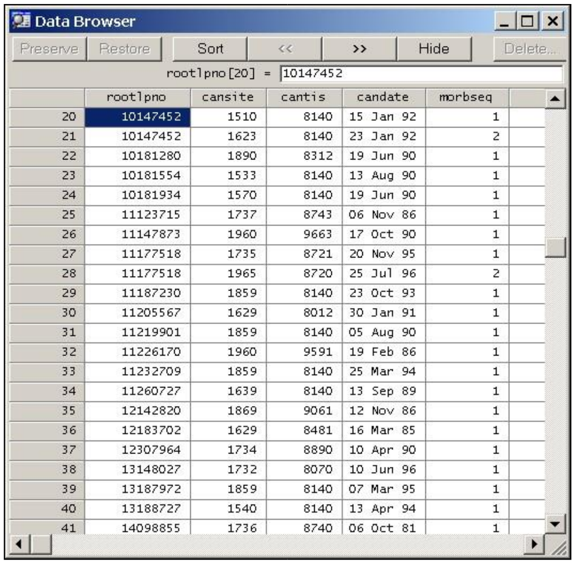
	rootlpno	cansite	cantis	candate	morbseq
20	10147452	1510	8140	15.01.92	1
21	10147452	1623	8140	23.01.92	2
22	10181280	1890	8312	19.06.90	1
23	10181554	1533	8140	13.08.90	1
24	10181934	1570	8140	19.06.90	1
25	11123715	1737	8743	06.11.86	1
26	11147873	1960	9663	17.10.90	1
27	11177518	1735	8721	20.11.95	1
28	11177518	1965	8720	25.07.96	2
29	11187230	1859	8140	23.10.93	1
30	11205567	1629	8012	30.01.91	1
31	11219901	1859	8140	05.08.90	1
32	11226170	1960	9591	19.02.86	1
33	11232709	1859	8140	25.03.94	1

SAS ↓



	ROOTLPNO	CANSITE	CANTIS	CANDATE	morbseq
20	10147452	1510	8140	15/01/1992	1
21	10147452	1623	8140	23/01/1992	2
22	10181280	1890	8312	19/06/1990	1
23	10181554	1533	8140	13/08/1990	1
24	10181934	1570	8140	19/06/1990	1
25	11123715	1737	8743	06/11/1986	1
26	11147873	1960	9663	17/10/1990	1
27	11177518	1735	8721	20/11/1995	1
28	11177518	1965	8720	25/07/1996	2
29	11187230	1859	8140	23/10/1993	1
30	11205567	1629	8012	30/01/1991	1
31	11219901	1859	8140	05/08/1990	1
32	11226170	1960	9591	19/02/1986	1
33	11232709	1859	8140	25/03/1994	1
34	11260727	1639	8140	13/09/1989	1

Stata →



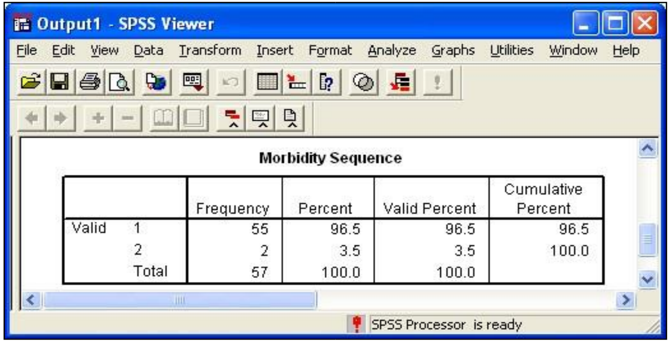
rootlpno[20] = 10147452

	rootlpno	cansite	cantis	candate	morbseq
20	10147452	1510	8140	15 Jan 92	1
21	10147452	1623	8140	23 Jan 92	2
22	10181280	1890	8312	19 Jun 90	1
23	10181554	1533	8140	13 Aug 90	1
24	10181934	1570	8140	19 Jun 90	1
25	11123715	1737	8743	06 Nov 86	1
26	11147873	1960	9663	17 Oct 90	1
27	11177518	1735	8721	20 Nov 95	1
28	11177518	1965	8720	25 Jul 96	2
29	11187230	1859	8140	23 Oct 93	1
30	11205567	1629	8012	30 Jan 91	1
31	11219901	1859	8140	05 Aug 90	1
32	11226170	1960	9591	19 Feb 86	1
33	11232709	1859	8140	25 Mar 94	1
34	11260727	1639	8140	13 Sep 89	1
35	12142820	1869	9061	12 Nov 86	1
36	12183702	1629	8481	16 Mar 85	1
37	12307964	1734	8890	10 Apr 90	1
38	13148027	1732	8070	10 Jun 96	1
39	13187972	1859	8140	07 Mar 95	1
40	13188727	1540	8140	13 Apr 94	1
41	14098855	1736	8740	06 Oct 81	1

10. Request a frequency distribution on morbseq. The results should show that there are 55 first-time and 2 second-time cancers in the file.

11. Now reconstruct this file such that each individual only has one record containing all of their available cancer data.

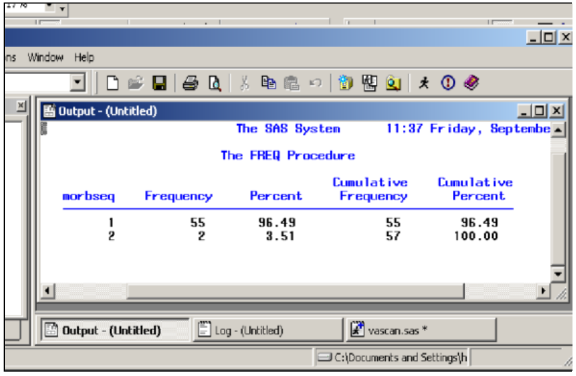
SPSS


Output1 - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help
SPSS Processor is ready

Morbidity Sequence

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	55	96.5	96.5	96.5
	2	2	3.5	3.5	100.0
	Total	57	100.0	100.0	

SAS



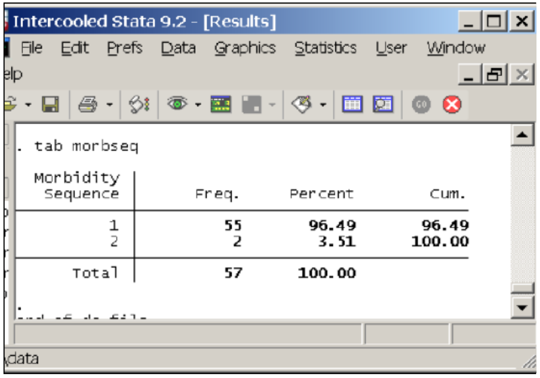
```
                    The SAS System        11:37 Friday, Septembe

                  The FREQ Procedure

                                          Cumulative    Cumulative
morbseq    Frequency     Percent          Frequency      Percent

      1          55        96.49                55          96.49
      2           2         3.51                57         100.00
```

Stata



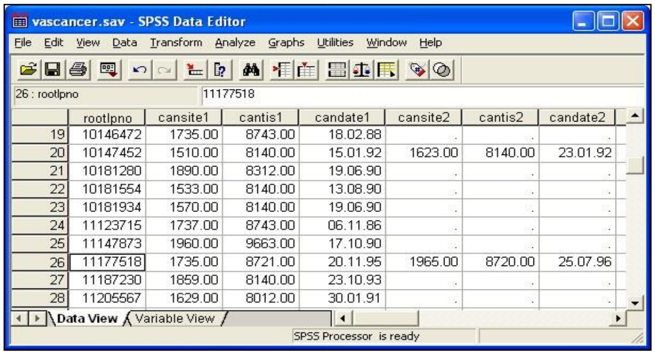
```
. tab morbseq

  Morbidity |
   Sequence |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         55       96.49       96.49
          2 |          2        3.51      100.00
------------+-----------------------------------
      Total |         57      100.00
```

Important note for everyone: Please follow the instructions below carefully and do not waste your time by trying to work out a better way to perform the task. The exercise seeks to demonstrate a principle by asking you to take a bad approach to the task – a method that you will never use again once you have acquired a few more skills over the days to come.

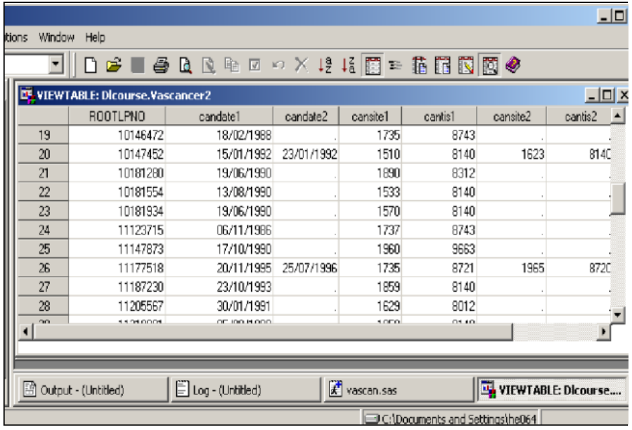
Compose and execute syntax that will create six new variables in the EOR loading area: cansite1, cantis1, candate1, cansite2, cantis2 and candate2. For records of first-time cancers, the syntax should assign the values of cansite, cantis and candate to cansite1, cantis1 and candate1 on the <u>same</u> record; while leaving cansite2, cantis2 and candate2 blank. For records of second-time cancers, the syntax should assign the values of cansite, cantis and candate <u>found on the previous record</u> to cansite1, cantis1 and candate1; while assigning the values of cansite, cantis and candate found on the <u>same</u> record to cansite2, cantis2 and candate2.

SPSS



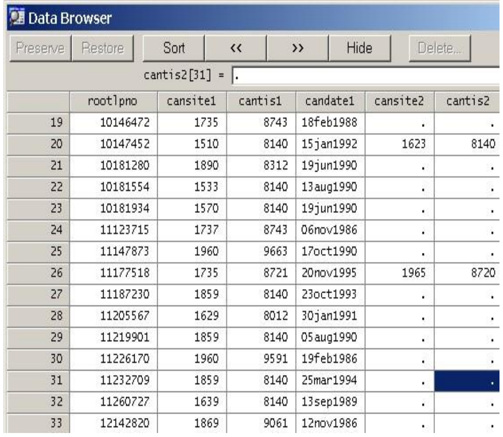
	rootlpno	cansite1	cantis1	candate1	cansite2	cantis2	candate2
19	10146472	1735.00	8743.00	18.02.88	.	.	.
20	10147452	1510.00	8140.00	15.01.92	1623.00	8140.00	23.01.92
21	10181280	1890.00	8312.00	19.06.90	.	.	.
22	10181554	1533.00	8140.00	13.08.90	.	.	.
23	10181934	1570.00	8140.00	19.06.90	.	.	.
24	11123715	1737.00	8743.00	06.11.86	.	.	.
25	11147873	1960.00	9663.00	17.10.90	.	.	.
26	11177518	1735.00	8721.00	20.11.95	1965.00	8720.00	25.07.96
27	11187230	1859.00	8140.00	23.10.93	.	.	.
28	11205567	1629.00	8012.00	30.01.91	.	.	.

SAS



	ROOTLPNO	candate1	candate2	cansite1	cantis1	cansite2	cantis2
19	10146472	18/02/1988	.	1735	8743	.	.
20	10147452	15/01/1992	23/01/1992	1510	8140	1623	8140
21	10181280	19/06/1990	.	1890	8312	.	.
22	10181554	13/08/1990	.	1533	8140	.	.
23	10181934	19/06/1990	.	1570	8140	.	.
24	11123715	06/11/1986	.	1737	8743	.	.
25	11147873	17/10/1990	.	1960	9663	.	.
26	11177518	20/11/1995	25/07/1996	1735	8721	1965	8720
27	11187230	23/10/1993	.	1859	8140	.	.
28	11205567	30/01/1991	.	1629	8012	.	.

Stata



cantis2[31] = .

	rootlpno	cansite1	cantis1	candate1	cansite2	cantis2
19	10146472	1735	8743	18feb1988	.	.
20	10147452	1510	8140	15jan1992	1623	8140
21	10181280	1890	8312	19jun1990	.	.
22	10181554	1533	8140	13aug1990	.	.
23	10181934	1570	8140	19jun1990	.	.
24	11123715	1737	8743	06nov1986	.	.
25	11147873	1960	9663	17oct1990	.	.
26	11177518	1735	8721	20nov1995	1965	8720
27	11187230	1859	8140	23oct1993	.	.
28	11205567	1629	8012	30jan1991	.	.
29	11219901	1859	8140	05aug1990	.	.
30	11226170	1960	9591	19feb1986	.	.
31	11232709	1859	8140	25mar1994	.	.
32	11260727	1639	8140	13sep1989	.	.
33	12142820	1869	9061	12nov1986	.	.

For the sake of aesthetics, you may delete the original cansite, cantis and candate variables and the morbseq variable from the file. You should also manually delete the original first-time cancer records for rootlpno 10147452 and rootlpno 11177518. The appearance of the data file should now be like that shown above with the two sets of cancer information.

As already emphasised, the methods you have just used in step 4 are poor practice and not the approach that you will ever use again once you have more skills. It should never be necessary to delete records manually in a problem of this type. The reason you are 'forced' to do it here is because for the individuals with two cancer records, the record receiving the aggregated information on both cancers was the second record and not the first record. This leaves us in the situation where we have no syntactical basis on which to define which first records are followed by a second record for the same individual and may thus be deleted. You had to look at the file visually to work this out and then delete manually. If the file was much bigger and there was the possibility of 1-5 cancer records per patient, this visual and manual approach would become quickly impracticable.

It is possible to write more generic syntax that copes with a file of any size and a larger, multiple number of cancers and to perform the task in a way that loads all of the information onto the EOR area of the first-time cancer record. This more generic syntax would thus enable all first-time records to be easily selected using morbseq = 1 for saving as a new file, rather than manually deleting the unwanted records. Methods for achieving this result involve either the use of an 'upside-down file' or a 'vector file' and you will learn about these later in the course.

12. Save the reconstructed data file using the file name vascancer2. You will need it for later work.

13. Finally, save your syntax written for exercise 1, as some of it will be used again. **Useful tip**: add documentation by way of non-executable comments or remarks to your syntax. Most computing languages allow this. Documentation on syntax will prove invaluable when you may wish to use it again several years later or if you wish to share your syntax with a colleague.

References

Borgmeier I, Holman CDJ. **Does vasectomy reversal protect against prostate cancer?** *Ann Epidemiol* 2004; 14: 748-749.

Cody R. **Longitudinal data techniques: looking across observations.** Available at: http://www2.sas.com/proceedings/sugi27/p015-27.pdf.