

12 Statistics, computers and calculators

12.1 Introduction

It must be admitted that many of the statistical procedures discussed in this book involve quite lengthy and tedious calculations, in the course of which human calculators are likely to make the occasional error. It is not surprising, then, that much statistical work is now carried out by means of computers. Space does not permit more than a brief discussion of this aspect of the use of computers in language study: readers are referred to the companion volume to this book (Butler, 1985) for a survey of computing in linguistic and literary studies.

Fortunately, researchers who wish to use a computer to relieve them of some of the drudgery of statistical calculations can often do so without having to learn to program the machine in one of the so-called 'high-level' computer languages such as BASIC, FORTRAN or PASCAL. Because statistical calculations are a common requirement of a wide range of users, 'package' programs have been developed, which allow users to specify their needs in terms that are not too far removed from ordinary English; the translation of these instructions into a form on which the computer can act directly is something that they need know nothing about. In this concluding chapter, the use of two widely available statistical packages, the Statistical Package for the Social Sciences (SPSS) and Minitab, will be illustrated, using some of the examples worked through manually in previous chapters. There then follows a brief review of the usefulness of the simple electronic calculator in statistical work.

12.2 The Statistical Package for the Social Sciences (SPSS)

SPSS is an extremely comprehensive statistical analysis package, supported on a wide range of computers. As the name suggests, it was designed primarily for use by social scientists, and the concept on which SPSS analyses are based, namely the 'case', reflects this orientation. A 'case' in a sociological study might represent a particular member of the human group under investigation, together with values for various attributes such as age, sex, social class and the like. Clearly, such a concept is directly applicable also to some kinds of sociolinguistic study. In other types of linguistic and literary investigation the 'case' concept is less clearly relevant; however, with a little thought the data can usually be expressed in a form suitable for application of the package.

SPSS will produce frequency distributions, bar charts and descriptive statistics (mean, median, mode, range, standard deviation, standard error). It will allow cross-tabulation of data in contingency tables, and performance of chi-square tests to assess the degree of association of variables. Parametric tests (the *t*-test for independent or correlated samples, analysis of variance) and a range of non-parametric tests (including the Mann-Whitney, Wilcoxon signed-ranks and sign tests) are available. Scattergrams can be produced, and Pearson or Spearman correlation coefficients calculated. More sophisticated analyses which are beyond the scope of this book can also be carried out. An overview of the facilities offered can be found in the introductory guide by Norušis (1982). In addition to the normal 'batch' mode version of SPSS, in which output from the analysis goes either to the lineprinter or to a named file, there is a version which allows the 'conversational' or 'interactive' use of the package, the results of the analysis being sent directly to the terminal at which the user is working.

As our first example of the use of SPSS, let us take the production of a frequency distribution table and descriptive statistics for the scores of class A in the language test situation discussed in chapters 2 and 3 (for data, see table 2.1). A suitable set of SPSS commands is shown in figure 12.1. The RUN NAME command simply gives a title to the analysis. The package is then informed that only one variable, called SCORE, is involved. It is also told that the input data are in free format; that is, they are strung out

along one or more lines, individual data items being separated by blanks. There are 30 cases (in this example, a case is just the score for one language learner). The VAR LABELS command specifies that the full title 'SCORE ON LANGUAGE TEST' is to be associated with the variable SCORE. The FREQUENCIES command says that it is the variable SCORE that is to be analysed. In the STATISTICS command, options 1, 3, 4, 5 and 9 are selected: these correspond to the mean, median, mode, standard deviation and range, respectively. The package is then instructed to read the input data, which follow the read command. The end of the data is signalled, and the FINISH command terminates the run. Figure 12.2 shows the frequency distribution table and statistics which are printed in response to these commands. Note that the frequency distribution table gives relative frequencies, cumulative frequencies and frequencies adjusted for any missing values (of which there are none in the present analysis), as well as the absolute frequencies. The values calculated for the descriptive statistics agree with those calculated in chapter 3.

As a second example of the usefulness of SPSS, consider the *t*-test used in section 7.2.1.2 to assess the significance of differences in the correctness of recall of sentences by two independent groups of subjects, tested under different conditions (for data, see table 7.2). The SPSS commands and data are shown in figure 12.3. The package is told that two variables, GROUP and RECALL, are involved: in the later VAR LABELS command these are associated with the full titles 'CONDITION OF TEST' and 'SENTENCES RECALLED CORRECTLY', respectively. Further, the VALUE LABELS command specifies that scores from Group 1 are identified in the data by the code 1, and scores from Group 2 by the

```

RUN NAME           LANGUAGE TEST SCORES, GROUP A
VARIABLE LIST      SCORE
INPUT FORMAT       FREEFIELD
N OF CASES         30
VAR LABELS         SCORE ON LANGUAGE TEST/
FREQUENCIES        GENERAL = SCORE
STATISTICS         1 3 4 5 9
READ INPUT DATA
15 12 11 18 15 15 9 19 14 13 11 12 18 15 16 14 16 17 15 17 13 14 13 15 17
19 17 18 16 14
END INPUT DATA
FINISH

```

Figure 12.1 SPSS commands and data for analysis of language test scores

SCORE ON LANGUAGE TEST		ABSOLUTE	RELATIVE	ADJUSTED	CUM
CATEGORY LABEL	CODE	FREQ	FREQ (PCT)	FREQ (PCT)	FREQ (PCT)
	9.	1	3.3	3.3	3.3
	11.	2	6.7	6.7	10.0
	12.	2	6.7	6.7	16.7
	13.	3	10.0	10.0	26.7
	14.	4	13.3	13.3	40.0
	15.	6	20.0	20.0	60.0
	16.	3	10.0	10.0	70.0
	17.	4	13.3	13.3	83.3
	18.	3	10.0	10.0	93.3
	19.	2	6.7	6.7	100.0
	TOTAL	30	100.0	100.0	
MEAN	14.933	MEDIAN	15.000	MODE	15.000
STD DEV	2.490	RANGE	10.000		
VALID CASES	30	MISSING CASES	0		

LANGUAGE TEST SCORES, GROUP A

Figure 12.2 SPSS output for analysis of language test scores

```
VARIABLE LIST  GROUP RECALL
INPUT FORMAT  FREEFIELD
N OF CASES    19
VAR LABELS    GROUP CONDITION OF TEST/
              RECALL SENTENCES RECALLED CORRECTLY/
VALUE LABELS  GROUP (1) CONDITION 1 (2) CONDITION 2
T-TEST       GROUPS=GROUP/VARIABLES=RECALL
READ INPUT DATA
1 18 1 15 1 13 1 17 1 14 1 8 1 10 1 11 1 7 1 17 2 13 2 14 2 12 2 6 2 11
2 13 2 17 2 16 2 5
END INPUT DATA
FINISH
```

Figure 12.3 SPSS commands and data for *t*-test on sentence recall

code 2. The INPUT FORMAT is specified as free, so that the data consist of a set of scores, each with its associated group coding (1 or 2), separated by blanks. The number of scores is given as 19 for the two groups taken together. A *t*-test is requested, with the groups for comparison being selected on the GROUP coding, and RECALL as the variable under test. The output (see figure 12.4) gives a number of pieces of information about the data. The number of cases, and the mean, standard deviation and standard

		T-TEST									
GROUP 1 - GROUP	EQ 1.	NUMBER	MEAN	STANDARD	STANDARD	F		POOLED VARIANCE ESTIMATE		SEPARATE VARIANCE ESTIMATE	
GROUP 2 - GROUP	EQ 2.	OF CASES		DEVIATION	ERROR	VALUE	2-TAIL	T	DEGREES OF	T	2-TAIL
VARIABLE				ERROR		VALUE	PROB.	VALUE	FREEDOM	VALUE	PROB.
RECALL											
GROUP 1		10	13.0000	3.887	1.229	1.10	0.883	0.61	17	0.61	0.552
GROUP 2		9	11.8889	4.076	1.359						

Figure 12.4 SPSS analysis output for *t*-test on sentence recall data

error for the variable under test (here, the number of sentences recalled correctly) are given for each group, and the F ratio computed to test for homogeneity of variance. The probability of obtaining the calculated F value or a greater value is high (0.883), so that we have no cause to abandon the homogeneity of variance assumption. Two t -values are calculated, one based on a pooled variance estimate, as in the method discussed in chapter 7, the other based on separate variance estimates for the two groups, a technique (not discussed in this book) for cases where the variances of the two groups differ considerably. In the present case, as we have seen, the variances are very similar, so it is not surprising that the t -values arrived at by the two methods are equal. The slight difference between the value given by SPSS (0.61) and that calculated in chapter 7 (0.60) is due simply to the effect of rounding error. As the t -value calculated has a probability of 0.551 of being reached, we have no grounds for rejecting the null hypothesis that the two groups are from distributions with the same mean.

As a third example, consider the use of the Wilcoxon signed-ranks test to assess the significance of differences in the numbers of mistakes made by ten subjects in translating two passages of English into French. This problem was discussed in section 8.3 (for data, see table 8.3). The SPSS commands and data are shown in figure 12.5. A name is given to the run, and the package is told that two variables, named PA and PB, are involved. The format of the data is specified as free, and there are ten cases (that is, ten pairs of scores, each corresponding to one subject). The labels PA and PB are to be associated with the names 'PASSAGE A' and 'PASSAGE B', for use in the labelling of the output. A Wilcoxon

```

RUN NAME          TRANSLATION ERRORS IN TWO PASSAGES OF ENGLISH TO FRENCH
VARIABLE LIST    PA PB
INPUT FORMAT     FREEFIELD
N OF CASES       10
VAR LABELS       PA PASSAGE A
                  PB PASSAGE B
NPAR TESTS       WILCOXON = PA WITH PB
READ INPUT DATA
8 10 7 6 4 4 2 5 4 7 10 11 17 15 3 6 2 3 11 14
END INPUT DATA
FINISH
    
```

Figure 12.5 SPSS commands and data for Wilcoxon test on translation errors

WILCOXON MATCHED-PAIRS SIGNED-RANKS TEST

PA WITH PB		PASSAGE A			
CASES	TIES	2 -RANKS MEAN	7 +RANKS MEAN	Z	2-TAILED P
10	1	3.25	5.50	-1.896	0.058

TRANSLATION ERRORS IN TWO PASSAGES OF ENGLISH TO FRENCH

Figure 12.6 SPSS analysis output for Wilcoxon test on translation errors

test of passage A against passage B is requested. The input data are read in, and the run terminated. The results are shown in figure 12.6. The package in fact calculates a z-score based on the formula given in section 8.3, even though the number of pairs is considerably smaller than the 20 or so normally considered necessary for testing using the z-score method. The probability of obtaining the calculated value or a more extreme one is given as 0.058, so that we cannot reject, at the 5 per cent level, the null hypothesis that the two sets of scores share the same distribution, although the value is very near to the critical region.

As a final example, let us take the problem, discussed in chapter 9, of testing the degree of association between sentence length (long, medium or short) and which of three novels the sentence occurs in (for data, see table 9.5). The SPSS commands and data are shown in figure 12.7. A name is given to the analysis, and the package is then informed that the data are presented in fixed format, rather than in the free format used in earlier examples. The frequency for a particular cell in table 9.5 is in positions 1 to 3 in the data line, a numerical coding for the length of sentence (see below) is in position 5, and a further coding representing the novel concerned is in position 7. The WEIGHT command instructs the package to treat the frequency in each cell as representing that number of individual cases, each of the same kind. The variable LENGTH is to be associated with the longer name 'LENGTH OF SENTENCE'. There are three codings for LENGTH: 1 represents a short sentence, 2 a medium-length sentence, and 3 a long sentence. Similarly, the codes 1, 2 and 3 are used to represent novels 1, 2 and 3, respectively. A cross-tabulation (that is, production of a contingency table) of the sentence length variable by the novel variable is requested. The STATISTICS command selects option 1,

```
RUN NAME          SENTENCE LENGTH IN THREE NOVELS
DATA LIST         FIXED/1  FREQ 1-3  LENGTH 5  NOVEL 7
WEIGHT           FREQ
VAR LABELS       LENGTH, LENGTH OF SENTENCE/
VALUE LABELS     LENGTH (1) SHORT (2) MEDIUM (3) LONG/
                  NOVEL (1) NOVEL 1 (2) NOVEL 2 (3) NOVEL 3/
CROSSTABS        TABLES = LENGTH BY NOVEL
STATISTICS       1
READ INPUT DATA
103 1 1
 82 1 2
110 1 3
281 2 1
262 2 2
276 2 3
116 3 1
145 3 2
124 3 3
END INPUT DATA
FINISH
```

Figure 12.7 SPSS commands and data for cross-tabulation and chi-square analysis of sentence length in three novels

which is in fact the chi-square test. The data are read in, and the analysis terminated. The contingency table and chi-square result produced are shown in figure 12.8; the slight difference between the chi-square value and that calculated in chapter 9 is due to rounding error. SPSS gives the probability of obtaining a chi-square value greater than or equal to the calculated value. Since this is 0.0857, we cannot reject, at the 5 per cent level, the null hypothesis of no association. This is, of course, the conclusion we reached in chapter 9.

12.3 The Minitab package

Minitab was designed at the Pennsylvania State University, as an all-purpose, flexible data manipulation package for users with little statistical or computational knowledge. Like SPSS, it is available on a wide range of computers. Minitab is perhaps rather easier to use than SPSS: it is based on a 'worksheet' of rows and columns of data, rather than on the 'case' principle, and the instructions to the package are rather simpler and more transparent than those for SPSS. Minitab has facilities for performing arithmetical operations (such as adding the contents of two columns), for editing, for

***** CROSSTABULATION OF *****
 LENGTH LENGTH OF SENTENCE BY NOVEL

		NOVEL			ROW TOTAL
		NOVEL 1	NOVEL 2	NOVEL 3	
LENGTH	COUNT				
	ROW PCT				
SHORT	COL PCT				
	TOT PCT				
		1.	2.	3.	
SHORT	1.	103	82	110	295
		34.9	27.8	37.3	19.7
		20.6	16.8	21.6	
		6.9	5.5	7.3	
MEDIUM	2.	281	262	276	819
		34.3	32.0	33.7	54.6
		56.2	53.6	54.1	
		18.7	17.5	18.4	
LONG	3.	116	145	124	385
		30.1	37.7	32.2	25.7
		23.2	29.7	24.3	
		7.7	9.7	8.3	
COLUMN TOTAL		500	489	510	1499
		33.4	32.6	34.0	100.0

CHISQUARE = 8.16640 WITH 4 DEGREES OF FREEDOM SIGNIFICANCE = 0.0857

SENTENCE LENGTH IN THREE NOVELS

Figure 12.8 SPSS analysis output for sentence length data

sorting data into rank order, and other commonly required types of manipulation. Its range of statistical facilities is wide, though not quite so comprehensive as that found in SPSS. For instance, the Wilcoxon signed-ranks test is not available, though the Mann-Whitney test is included. A simple guide to Minitab can be found in Ryan, Joiner and Ryan (1976).

Figure 12.9 shows a set of Minitab commands and data for the production of a frequency distribution and descriptive statistics for the language test data analysed in our first SPSS example. The command SET THE FOLLOWING INTO C1 simply reads the set of 30 scores, supplied in free format in a list following the instruction, and sets them into a column which can later be referred to as C1. The remaining commands request the mean, standard deviation and median (the mode and range are not available), and finally a histogram of the data. The output is shown in figure

```

SET THE FOLLOWING INTO C1
15 12 11 18 15 15 9 19 14 13 11 12 18 15 16 14 16 17 15 17 13 14 13 15 17
19 17 18 16 14
AVERAGE OF C1
STANDARD DEVIATION OF C1
MEDIAN OF C1
HISTOGRAM OF C1
STOP
    
```

Figure 12.9 Minitab commands and data for analysis of language test scores

```

-- SET THE FOLLOWING INTO C1
COLUMN      C1
COUNT      30
           15.          12.          11.          18. ...

-- AVERAGE OF C1
AVERAGE = 14.933

-- STANDARD DEVIATION OF C1
ST. DEV. = 2.4904

-- MEDIAN OF C1
MEDIAN = 15.000

-- HISTOGRAM OF C1

C1

MIDDLE OF   NUMBER OF
INTERVAL    OBSERVATIONS
  9.         1      *
 10.         0
 11.         2     **
 12.         2     **
 13.         3    ***
 14.         4   ****
 15.         6   *****
 16.         3    ***
 17.         4   ****
 18.         3    ***
 19.         2     **

-- STOP
    
```

Figure 12.10 Minitab analysis output for language test data

12.10. Note that Minitab produces a sideways-on histogram consisting of a number of asterisks proportional to the frequency in each class.

In figure 12.11 is shown a file of commands and data for performing the *t*-test on sentence recall data, discussed earlier in relation to SPSS. The two sets of scores are set into two columns,

```

SET INTO C1
18 15 13 17 14 8 10 11 7 17
SET INTO C2
13 14 12 6 11 13 17 16 5
POOLED T FOR DATA IN C1 AND C2
STOP

```

Figure 12.11 Minitab commands and data for *t*-test on sentence recall

```

-- SET INTO C1
COLUMN      C1
COUNT      10
           18.          15.          13.          17. ...

-- SET INTO C2
COLUMN      C2
COUNT      9
           13.          14.          12.          6. ...

-- POOLED T FOR DATA IN C1 AND C2
C1          N = 10          MEAN = 13.000          ST. DEV. = 3.89
C2          N = 9          MEAN = 11.889          ST. DEV. = 4.08

DEGREES OF FREEDOM = 17

A 95.00 PERCENT C.I. FOR MU1-MU2 IS ( -2.7452, 4.9674)

TEST OF MU1 = MU2 VS. MU1 N.E. MU2
T = 0.608
THE TEST IS SIGNIFICANT AT 0.5512
CANNOT REJECT AT ALPHA = 0.05

-- STOP

```

Figure 12.12 Minitab analysis output for *t*-test on sentence recall

C1 and C2, and a *t*-test based on a pooled variance estimate is requested. The output (figure 12.12) gives the number of scores, mean and standard deviation, for each group, the number of degrees of freedom involved, the 95 per cent confidence interval for the difference between the means, the value of *t*, the probability of attaining such a value, and the conclusion to be drawn in a test at the 5 per cent level.

Finally, let us consider how, using Minitab, we might construct a contingency table and calculate chi-square for the data on sentence length in three novels, discussed earlier. The file of commands and data is shown in figure 12.13. The package is instructed to read the data into a three-column table, and then to perform

```

READ THE TABLE INTO C1, C2, C3
103  82  110
281 262 276
116 145 124
CHISQUARE ANALYSIS ON TABLE IN C1, C2, C3
STOP
    
```

Figure 12.13 Minitab commands and data for cross-tabulation and chi-square analysis of sentence length in three novels

```

-- READ THE TABLE INTO C1, C2, C3
COLUMN          C1          C2          C3
COUNT          3          3          3
ROW
  1             103.        82.        110.
  2             281.        262.       276.
  3             116.        145.       124.

-- CHISQUARE ANALYSIS ON TABLE IN C1, C2, C3

EXPECTED FREQUENCIES ARE PRINTED BELOW OBSERVED FREQUENCIES

```

	C1	C2	C3	TOTALS
1	103 98.4	82 96.2	110 100.4	295
2	281 273.2	262 267.2	276 278.6	819
3	116 128.4	145 125.6	124 131.0	385
TOTALS	500	489	510	1499

```

TOTAL CHI SQUARE =

    0.22 + 2.11 + 0.92 +
    0.22 + 0.10 + 0.03 +
    1.20 + 3.00 + 0.37 +

    = 8.17

DEGREES OF FREEDOM = (3 - 1) X (3 - 1) = 4

-- STOP
    
```

Figure 12.14 Minitab analysis output for sentence length data

a chi-square analysis. The output (figure 12.14) shows the contingency table, the contribution to chi-square made by each cell in the table, and the total chi-square, but, curiously, no figure for the probability associated with this value, which must be looked up in tables by the user.

12.4 The use of electronic calculators in statistical work

The student or researcher who has no access to a statistical package, and who is himself unable to program the computer to perform the statistical analyses he requires, can still benefit considerably from electronic aids. The wide range of quite inexpensive pocket

Table 12.1 The use of a calculator to compute a mean and standard deviation

<i>Figure entered</i>	<i>Function button(s) depressed</i>	<i>Screen display after depression of function button</i>
15	$\Sigma +$	1.
12	$\Sigma +$	2.
11	$\Sigma +$	3.
18	$\Sigma +$	4.
15	$\Sigma +$	5.
15	$\Sigma +$	6.
9	$\Sigma +$	7.
19	$\Sigma +$	8.
14	$\Sigma +$	9.
13	$\Sigma +$	10.
11	$\Sigma +$	11.
12	$\Sigma +$	12.
18	$\Sigma +$	13.
15	$\Sigma +$	14.
16	$\Sigma +$	15.
14	$\Sigma +$	16.
16	$\Sigma +$	17.
17	$\Sigma +$	18.
15	$\Sigma +$	19.
17	$\Sigma +$	20.
13	$\Sigma +$	21.
14	$\Sigma +$	22.
13	$\Sigma +$	23.
15	$\Sigma +$	24.
17	$\Sigma +$	25.
19	$\Sigma +$	26.
17	$\Sigma +$	27.
18	$\Sigma +$	28.
16	$\Sigma +$	29.
14	$\Sigma +$	30.
	2nd mean	14.933 333
	2nd st. dev.	2.490 441 5

calculators now available includes models with built-in statistical functions, or even with limited facilities for programming. Even models which do not have built-in functions will usually allow intermediate storage of results in the calculator's memory, thereby facilitating, for example, the accumulation of sums of squares, as well as sums of individual scores, in the calculation of a standard deviation.

One popular and inexpensive model of calculator has facilities for the automatic computation of the mean, standard deviation and variance of a set of data, also the Pearson correlation coefficient for two sets of scores. In table 12.1 is shown the sequence of operations for calculating the mean and standard deviation of the 30 language test scores processed earlier by means of SPSS and Minitab.

12.5 A final caveat

There is, of course, every reason why we should take advantage of computers or pocket calculators to reduce the tedium and the likelihood of error involved in calculating statistics from first principles. However, it cannot be too strongly emphasised that it is essential to understand the principles on which the calculation and use of the statistics are based. The competent scholar will have the knowledge to decide when a particular descriptive statistic or test is the most appropriate one to use, and will be thoroughly conversant with the limitations, as well as the advantages, of the procedures he adopts. It is hoped that this book will help to increase the level of awareness and understanding of statistics among that varied group of scholars whose interest is in the analysis of language.