CHAPTER 14
Meaningful Data

What do you do when presented with a large amount of numerical data? For instance, here's a set of numbers which could be the result of some experiment, results of some examination, etc.

```
23  67  89 45 40 10   5
19  99  40 23  9 11  21
34  34  56 41 42 27  80
```

You can gain a great deal of information about the data by looking at such descriptive measures as the mean, the variance and the standard deviation. But first the data has to be entered into our computer.

Such data could be stored on your computer by using an array X(1), so that X(0) = 23, X(1) = 67, etc. It could be entered interactively by a simple program. This is illustrated with the programs Data Entry I and Data Entry II.

Listing 14.1
LIST

```
  200 REM Data entry I
  210 MODE 6:VDU 19,0,4,0,0,0,0:PRINT '
TAB(15);"Data entry"'
  220 PRINT "This allows you to enter so
me numerical data (at least 2)."'
  230 PRINT "How much data do you want t
o enter?"
  240 REPEAT
  250  INPUT '"Number? " N
  260  IF N<2 OR N<>INT(N) THEN PRINT '"
Be reasonable!"
  270 UNTIL N>1 AND N=INT(N)
  280 N=N-1:DIM X(N)
  290 PRINT '"Now enter the ";N+1;" item
 of data."'
```

```
  300 FOR I=0 TO N
  310  PRINT "Data number ";I+1;" ";:INP
UT X(I)
  320 NEXT
  330 PRINT CHR$(7) ' TAB(9);"Press Y to
 continue. ";
  340 REPEAT:UNTIL GET$="Y"
  350 PRINT
```

RUN

```
                Data entry

This allows you to enter some numerical
data (at least 2).

How much data do you want to enter?

Number? 1.2

Be reasonable!

Number? 3

Now enter the 3 item of data.

Data number 1 ?1.2
Data number 2 ?3.67
Data number 3 ?1.99

          Press Y to continue.
```

Listing 14.2
LIST

```
  200 REM Data entry II
  210 MODE 6:VDU 19,0,4,0,0,0,0:PRINT '
TAB(15);"Data entry"'
  220 PRINT "This allows you to enter so
```

```
me numerical data (at least 2)."'
  230 PRINT "Enter your data item by ite
m."'
  240 M=100:DIM X(M)
  250 FOR I=0 TO M
  260  IF I>1 THEN PRINT TAB(9+LEN(STR$(
I+1)));"Type -99999 to end."
  270  PRINT "Data number ";I+1;" ";:INP
UT X(I)
  280  IF X(I)=-99999 AND I>1 THEN N=I-1
:I=M
  290  IF X(I)=-99999 AND I<2 THEN PRINT
 TAB(9);"Too early to end.":I=I-1
  300 NEXT
  310 PRINT CHR$(7) ' TAB(9);"Press Y to
 continue. ";
  320 REPEAT:UNTIL GET$="Y"
```

Data Entry I asks, at the start, for the amount of data to be entered. The array X(I) is then dimensioned accordingly. On the other hand Data Entry II assumes that there are no more than 100 numbers to be entered. You enter the data when requested, to stop you enter -99999. If necessary, the number M in line 240 may be changed from 100 to some other number.

This chapter contains several short programs which may be added to the Data Entry programs. The final is a useful program which will help analyse your data.

The mean

The mean or average is an important statistical measure. It is obtained by adding all the numbers together and dividing the sum by the number of numbers.

$$\text{mean} = \frac{\text{sum of data}}{\text{number of data}}$$

If the numbers are stored in the array X(I) for I = 0 to N -1 then the mean XM may be calculated using the following program lines.

```
X = 0
FOR I = 0 TO N - 1 : X = X + X(I) : NEXT
XM = X/N
```

The program Mean is a short program that calculates the mean and prints it out. You may incorporate it with one of the Data Entry programs.

Listing 14.3
LIST

```
  400 REM Mean of data
  410 CLS:PRINT ' TAB(13);"Data analysis
"'
  420 PRINT "Number of data items = ";N+
1
  430 X=0:FOR I=0 TO N:X=X+X(I):NEXT:XM=
X/(N+1)
  440 PRINT '"Mean = ";XM
```

Max, min and spread

It's often useful to know the maximum and minimum values of data. A simple search may be made by your computer to find these. The following program is a short program which performs this function. In addition, the range or spread of the data is calculated. This is simply the difference between the largest and smallest numbers in the data.

Listing 14.4
LIST

```
  500 REM Max. Min and spread of data
  510 MAX=-1E37:MIN=1E37
  520 FOR I=0 TO N
  530  IF X(I)>MAX THEN MAX=X(I)
  540  IF X(I)<MIN THEN MIN=X(I)
  550 NEXT
  560 PRINT '"Minimum value = ";MIN
  570 PRINT "Maximum value = ";MAX
  580 PRINT "The spread is = ";MAX-MIN
```

You may want your data sorted into increasing (or decreasing) order. Several methods (such as bubble sort, quick short, shell sort, etc.) are available. Details are not given here.

Standard deviation and variance

The mean is a simple, useful and powerful tool. But it does not tell us all we need to know. For instance, look at the following sets of data:

DATA for X(I) 20, 21, 20, 19
DATA for Y(I) 38, 26, 14, 2

The values of the means XM and YM are both 20. However there is such greater variation in the data for Y(I) than for X(I). This variation may be measured by using the standard deviation of the data.

The *standard deviation* of a set of data is given by the following formula.

$$\text{standard deviation} = \text{SQR}\left(\frac{\text{sum (difference between data and mean)}^2}{\text{number of data - 1}}\right)$$

The variance is the square of the standard deviation.

The procedure for calculating the standard deviation of the data stored in the array X(I) is as follows.

1. Calculate the mean XM
2. Find the deviations from the mean, that is, the values X(I) - XM.
3. Square each deviation, that is, find the values (X(I) - XM)2.
4. Sum the squares of the deviations.
5. Divide by the number of terms less 1. This gives the variance XV of the data.
6. Take the square root. This is the standard deviation XD of the data.

The standard deviation provides an idea of how much the data is dispersed or spread out around the mean. Look at the following examples again.

DATA for X(I) 20, 21, 20, 19
DATA for Y(I) 38, 26, 14, 2

The standard deviations are given by the following calculations.

```
XD = SQR((0*0 + 1*1 + 0*0 + (-1)*(-1))/3)
   = SQR(2/3)
   = 0.816496581
YD = SQR((18*18 + 6*6 + (-6)*(-6) + (-18)*(-18))/3)
   = SQR(240)
   = 15.4919334
```

This certainly reflects the difference in the spread of the data.

The next short program is for calculating the standard deviation of the data stored in array X(I).

Listing 14.5

```
  600 REM Standard deviation
  610 X=0:FOR I=0 TO N:Y=X(I)-XM:X=X+Y*Y
:NEXT
  620 XD=SQR(X/N)
  630 PRINT '"Standard deviation = ";XD
```

Confidence intervals

The standard deviation is useful because it indicates to what extent the data is spread about the mean. In many mass manufacturing processes the product produced varies slightly in size or quality or length, etc. We refer to the items we are measuring as the population. The variation in the population is often normally distributed.

Statisticians have found that the normal curve or normal distribution approximates a large number of real-life data. If the amount of data is large (more than about 30) then it is often assumed, for calculations, that the population is normally distributed, even though this may not be so.

Roughly speaking a population is normally distributed if it is symmetrically spread about the mean; with most of the population at the mean and very little far away from the mean.

More precisely, in a normally distributed population about 68% of the population lies within 1 standard deviation from the mean, and about 96% lies within 2 standard deviations from the mean. More generally, the next table lists the percentages associated with various multiples of the standard deviation.

% of population	multiple of standard deviation
50%	0.6745
68.27%	1
80%	1.28
90%	1.645
95%	1.96
95.45%	2
99%	2.575
99.73%	3

The above table shows that we would expect 95% of a population to be within 1.96 times the standard deviation from the mean. In other words 95% of the population lies within the range

XM - 1.96*XD to XM + 1.96*XD

where XM is the mean and XD the standard deviation. We refer to this interval as the 95% confidence interval for the population. Similarly, the 99% confidence interval for the population is from

XM - 2.575*XD to XM + 2.575*XD

Here is an illustration of how confidence intervals may be useful. You suspect that a grocer is selling incompletely filled 2 litre bottles of

lemonade. You buy ten bottles and measure their contents carefully. The results in litres are:

2.001, 2.040, 2.020, 2.000, 2.015
2.006, 2.005, 2.031, 2.008, 2.018

All the bottles contain 2 litres or more. But, let's calculate the mean and standard deviation of this data. The results are:

XM = 2.0144
XD = 0.0132

Now, assuming that our sample came from a normally distributed population we can calculate some confidence intervals. The 95% confidence interval for the population is from

2.0144 - 1.96*0.0132 to 2.0144 + 1.96*0.0132

that is, from
1.989 to 2.040

Thus we expect 95% of the bottles to contain between 1.989 and 2.040 litres. This means that we expect 2.5% would contain more than 2.040 litres, while 2.5% would contain less than 1.989 litres. We could also work out the 90% confidence interval. This is from

2.0144 - 1.645*0.0132 to 2.0144 - 1.645*0.0132

or, from
1.993 to 2.036

Thus 90% of the bottles are expected to have between 1.993 and 2.036 litres. It follows that at least 5% of the lemonade bottles would contain less than the required 2 litres. (Equally, at least 5% contain more than 2.036 litres.)

In the example just given many assumptions have been made and the conclusions reached would be insufficient to lead to legal proceedings.

The mean of our population was calculated from a sample. How do we know that this is the actual mean of the population? The mean may vary with the sample taken. But, we can estimate how far our sample mean is from the real mean by using standard deviation. We can say with 95% confidence, that the mean is from

XM - 1.96*XD/SQR(N - 1) to XM + 1.96*XD/SQR(N - 1)

where XM is the mean calculated from the sample of size N. We call this the 95% confidence interval for the mean. The 99% confidence interval for the mean is given by

XM - 2.575*XD/SQR(N - 1) to XM + 2.575*XD/SQR(N - 1)

Strictly speaking these calculations are valid if the sample size is large (say greater than 30). For smaller samples we should be using what is called the *student's t distribution* instead of the normal distribution. But this is beyond the scope of this book.

Don't take the confidence intervals too seriously and don't confuse the two types of confidence interval.

The next program calculates 95% and 99% confidence intervals for the population and mean.

Listing 14.6
LIST

```
  700 REM Confidence intervals
  710 C1=1.96*XD:M1=C1/SQR(N):C2=2.575*X
D:M2=C2/SQR(N)
  720 PRINT '"95% Confidence intervals:"
  730 PRINT "Pop. from ";XM-C1;" to ";XM
+C1
  740 PRINT "Mean from ";XM-M1;" to ";XM
+M1
  750 PRINT '"99% Confidence intervals:"
  760 PRINT "Pop. from ";XM-C2;" to ";XM
+C2
  770 PRINT "Mean from ";XM-M2;" to ";XM
+M2
```

Finishing touches

The programs in this chapter may be linked together to produce a useful program for analysing data. The following program provides the necessary linking parts.

Listing 14.7
LIST

```
   10 REM Data analysis
```

```
   20 MODE 6:VDU 19,0,4,0,0,0,0:PRINT '
TAB(13);"Data analysis"'
   30 PRINT "This program allows you to
enter data   and analyse it."'
   40 PRINT "After the data has been ent
ered you willbe provided with the follow
ing informa- tion."'
   50 PRINT "1. Mean of data."'
   60 PRINT "2. Maximum, minimum and spr
ead of data."'
   70 PRINT "3. Standard deviation."'
   80 PRINT "4. 95% and 99% confidence i
ntervals."'
   90 PRINT ' TAB(9);"Press Y to continu
e. ";
  100 REPEAT:UNTIL GET$="Y"
```

Listing 14.8
LIST

```
  900 REM Ending
  910 PRINT ' CHR$(7) TAB(10);"Another g
o? Y or N ";
  920 REPEAT:G$=GET$:UNTIL G$="Y" OR G$=
"N"
  930 IF G$="Y" THEN RUN
  940 CLS:PRINT '"Bye for now.":END
```

The final result is the program given below.

Listing 14.9
LIST

```
   10 REM Data analysis
   20 MODE 6:VDU 19,0,4,0,0,0,0:PRINT '
TAB(13);"Data analysis"'
   30 PRINT "This program allows you to
enter data   and analyse it."'
   40 PRINT "After the data has been ent
ered you willbe provided with the follow
ing informa- tion."'
```

```
   50 PRINT "1. Mean of data."'
   60 PRINT "2. Maximum, minimum and spr
ead of data."'
   70 PRINT "3. Standard deviation."'
   80 PRINT "4. 95% and 99% confidence i
ntervals."'
   90 PRINT ' TAB(9);"Press Y to continu
e. ";
  100 REPEAT:UNTIL GET$="Y"
  200 REM Data entry II
  210 MODE 6:VDU 19,0,4,0,0,0,0:PRINT '
TAB(15);"Data entry"'
  220 PRINT "This allows you to enter so
me numerical data (at least 2)."'
  230 PRINT "Enter your data item by ite
m."'
  240 M=100:DIM X(M)
  250 FOR I=0 TO M
  260 IF I>1 THEN PRINT TAB(9+LEN(STR$(I
+1)));"Type -99999 to end."
  270 PRINT "Data number ";I+1;" ";:INPU
T X(I)
  280 IF X(I)=-99999 AND I>1 THEN N=I-1:
I=M
  290 IF X(I)=-99999 AND I<2 THEN PRINT
TAB(9);"Too early to end.":I=I-1
  300 NEXT
  310 PRINT CHR$(7) ' TAB(9);"Press Y to
 continue. ";
  320 REPEAT:UNTIL GET$="Y"
  400 REM Mean of data
  410 CLS:PRINT ' TAB(13);"Data analysis
"'
  420 PRINT "Number of data items = ";N+
1
  430 X=0:FOR I=0 TO N:X=X+X(I):NEXT:XM=
X/(N+1)
  440 PRINT '"Mean = ";XM
```

```
  500 REM Max. Min and spread of data
  510 MAX=-1E37:MIN=1E37
  520 FOR I=0 TO N
  530 IF X(I)>MAX THEN MAX=X(I)
  540 IF X(I)<MIN THEN MIN=X(I)
  550 NEXT
  560 PRINT '"Minimum value = ";MIN
  570 PRINT "Maximum value = ";MAX
  580 PRINT "The spread is = ";MAX-MIN
  600 REM Standard deviation
  610 X=0:FOR I=0 TO N:Y=X(I)-XM:X=X+Y*Y
:NEXT
  620 XD=SQR(X/N)
  630 PRINT '"Standard deviation = ";XD
  700 REM Confidence intervals
  710 C1=1.96*XD:M1=C1/SQR(N):C2=2.575*X
D:M2=C2/SQR(N)
  720  PRINT '"95% Confidence intervals:
"
  730 PRINT "Pop. from ";XM-C1;" to ";XM
+C1
  740 PRINT "Mean from ";XM-M1;" to ";XM
+M1
  750 PRINT '"99% Confidence intervals:"
  760 PRINT "Pop. from ";XM-C2;" to ";XM
+C2
  770 PRINT "Mean from ";XM-M2;" to ";XM
+M2
  900 REM Ending
  910 PRINT ' CHR$(7) TAB(10);"Another g
o? Y or N ";
  920 REPEAT:G$=GET$:UNTIL G$="Y" OR G$=
"N"
  930 IF G$="Y" THEN RUN
  940 CLS:PRINT '"Bye for now.":END
```

RUN

```
              Data analysis

This program allows you to enter data
and analyse it.

After the data has been entered you will
be provided with the following informa-
tion.

1. Mean of data.

2. Maximum, minimum and spread of data.

3. Standard deviation.

4. 95% and 99% confidence intervals.


          Press Y to continue.
                Data entry

This allows you to enter some numerical
data (at least 2).

Enter your data item by item.

Data number 1 ?2.001
Data number 2 ?2.040
           Type -99999 to end.
Data number 3 ?2.020
           Type -99999 to end.
Data number 4 ?2.000
           Type -99999 to end.
Data number 5 ?2.015
           Type -99999 to end.
Data number 6 ?2.006
```

```
          Type -99999 to end.
Data number 7 ?2.005
          Type -99999 to end.
Data number 8 ?2.031
          Type -99999 to end.
Data number 9 ?2.008
           Type -99999 to end.
Data number 10 ?2.018
           Type -99999 to end.
Data number 11 ?-99999

         Press Y to continue.

             Data analysis

Number of data items = 10

Mean = 2.0144

Minimum value = 2
Maximum value = 2.04
The spread is = 4.000000004E-2

Standard deviation = 1.317573529E-2

95% Confidence intervals:
Pop. from 1.988575559 to 2.040224441
Mean from 2.005791852 to 2.023008147

99% Confidence intervals:
Pop. from 1.980472481 to 2.048327518
Mean from 2.003090827 to 2.025709173

         Another go? Y or N
```