April 1996

Teaching with J - An Example from Statistics

Keith Smillie
Department of Computing Science
University of Alberta
Edmonton, Alberta T6G 2H1
smillie@cs.ualberta.ca

J is used in the presentation of some elementary concepts given in an introductory statistics text. The emphasis is on the development of the statistical material rather than on **J** with the language being introduced and discussed only as required for an orderly presentation of the topics in the text.

Introduction

Writing in 1991 on the 25th anniversary of APL, Kenneth Iverson remarked that "... Although APL has been exploited mostly in commercial programming, I continue to believe that its most important use remains to be exploited: as a simple, precise, executable notation for the teaching of a wide range of subjects." **J**, a modern dialect of APL, may be regarded similarly as a notation that will simplify and extend the computational aspects of any presentation and thus help give a greater understanding of the subject under discussion.

Because of the success of APL in a wide variety of commercial, industrial and scientific applications and also because of the enthusiasm with which the language has been used by its devotees for over twenty-five years, the original intent of the language has often been overlooked or even forgotten. Now that **J** is attracting increasing attention both within and outside the APL community, the time may be appropriate to emphasize the use of APL and **J** as a tool for exposition in the teaching of a specific topic.

To illustrate this theme we have used **J** to develop most of the computations in *Understanding Data* by Peter Sprent. Statistical topics discussed were taken from such areas as the tabular organization of data, stem-and-leaf diagrams, frequency distributions, data summaries, graphical presentation, off-the-cuff exploratory calculations, ranking methods, index numbers, simulation and sampling, regression analysis, and smoothing methods. Throughout this work the emphasis was on the presentation of statistical ideas with **J** being introduced as unobtrusively as possible as an aid to their understanding. Such an approach may be contrasted with a discussion of **J** with examples chosen from statistics to illustrate language concepts, an approach we like to think is still respectable.

This work has been influenced by a study of *Teach Yourself Business Japanese* by Michael Jenkins and Lynne Strugnell which is in the well-known and very successful "Teach Yourself Books" series. After an introductory chapter on Japanese syllabics, each of the subsequent chapters begins with an installment of a continuing story in Japanese followed by a list of new vocabulary, notes, further examples of the vocabulary and grammar just introduced, exercises, and a short essay in English on some aspect of Japanese culture or business practice. The point to emphasize here is that the Japanese language - vocabulary, grammar, style and polite usage - is introduced in the context of the continuing story of a Mr. Lloyd of the Overseas Planning Department of a British sporting goods firm and his relations both in Japan and in Britain with the company's Japanese representatives. (We learn, for example, phrases indispensable for conducting business in Japan such as *Biiru ni shimasho ka* "Shall we have a beer?" and *Biiru o mo ip-pon kudasai* "Another beer, please.")

Of course, Kenneth Iverson has advocated for many years the similarities between the teaching of APL, and now **J**, and the teaching of a natural language. An early paper on this subject is Iverson (1980). As with learning Japanese or any other natural language so in the present paper the **J** language is introduced and used in a realistic setting.

A detailed discussion of the material summarized in this paper may be found in Smillie (1996), the structure of which has been influenced by the Japanese text referred to above. In the present paper we shall restrict the examples to a few taken from the tabular presentation of data, summary statistics, exploratory calculations, and sampling methods.

Notation

The introductory remarks on **J** may be restricted to a few simple examples illustrating valence and precedence of verbs, defined verbs, adverbs and conjunctions, and

possibly forks. The following are the examples given in the present study:

```
   3 + 5              Plus
8
   3 - 5              Minus
_2
   2 * 3              Times
6
   15 % 6             Divided by
2.5
   % 2.5              Reciprocal
0.4
   2 * 3 + 4          Precedence
14
   (2 * 3) + 4
10
   % 15 % 6           Ambivalence
0.4
   +/1 2 3 4          Sum
10
   w=. 2.3 5 3.5 6
   +/w                Sum
16.8
   #w                 Tally
4
   (+/w) % #w         Arithmetic mean
4.2
   (+/ % #) w         Fork
4.2
   am=. +/ % #        Defined verb for
   am w                mean
4.2
   +/ % # w           Not the mean!
0.25
   w < 4              Less than
1 0 1 0
   +/w < 4            Number of items
2                      less than 4
   15 (<. @ %) 6      Floor atop divided by
2
   div=. <. @ %       Integer division
   15 div 6
2
   (div&10) 123       Bond (Integer division
12                          by 10)
      10 | 123        Residue
3
```

At the end of the statistical presentation in each of the sections of the work summarized here the verbs, adverbs and conjunctions that have been introduced are given in a box followed by a brief discussion. As an illustration the following box gives the parts of speech introduced in this section and shows, for example, that **%** represents the monadic verb *reciprocal* and the dyadic verb *divided by*, and that **/** is the adverb *insert* and **@** is the conjunction *atop*.

+	· *Plus*
-	· *Minus*
*****	· *Times*
%	*Reciprocal* · *Divided By*
#	*Tally* ·
<	· *Less than*
<.	*Floor* ·
|	· *Residue*
/	***Insert*** ·
@	**Atop**
&	**Bond**

Data organization

Data used in several of the examples in Sprent represent a company's computer downtime in minutes for May 1984 and appear in the text as Table 1.2. These data are given in **J** as table **T12** with **95** rows and **5** columns where the rows represent individual breakdowns and the columns represent date, time of breakdown, equipment involved, and duration in minutes, respectively. The type of equipment is coded as follows: **0** No breakdown, **1** CPU, **2** Disk drive, **3** Graph plotter, **4** Printer, **5** Service and **6** Tape reader. The first five rows of the table, which is given in the Appendix, are

```
1  842 4  49
2 1035 2  18
2 1529 1 123
2 1735 1   3
3    0 0   0
```

which show, for example, that on May 1 there was a Printer breakdown of 49 minutes starting at 8:42, on May 2 there were three breakdowns the first being a Disk drive breakdown of 18 minutes starting at 10:35, and on May 3 there were no breakdowns.

Individual columns of this table may be easily selected, and

```
X=. 3 {"1 T12
```

gives the durations

```
49 18 123 3 0 87 21 7 24 11 ...
```

in the last column and

```
E=. 2 {"1 T12
```

gives the type of failures

```
4 2 1 1 0 6 4 4 4 4 ...
```

in the third column. Since a day without an equipment failure is represented by an item in **E** equal to **0**, the corresponding items of **X** must be removed. This may be accomplished by

```
D=. (E > 0) # X
```

so that **D** has the value

```
49 18 123 3 87 21 7 24 11 19 ...
```

.

The items of **D** may be displayed conveniently as a table with **10** rows and **9** columns by the expression **10 9$D** which has the value

```
 49  18 123   3  87  21   7 24 11
 19 243  38  11  18  32 101  6 32
 27  41   3  25  19 242 122 15  7
  6   2  18  42  24   7   9 14 10
 30  23 141   7 102  83  29 11 15
  7   6   8 121   4  23  42 18 31
 25  42   6 339  19  27  12 11  3
  5   9   3  22  17  20   5  3 18
142  51  18  41   9   9  27 14 33
128  34  18 232 179 143 181  6 14
```

This table corresponds to Table 1.2 in Sprent.

Some very simple **J** expressions may be used to give some information about these data. For example, the number of breakdowns is **#D** or **90**, the shortest breakdown is **<./D** or **2** minutes and the longest is **>./D** or **339** minutes, and the average breakdown is **am D** or **44.9** minutes. The expression **D > 60** is equal to

```
0 0 1 0 1 0 0 0 0 0 ...
```

where the **1**s correspond to breakdowns greater than one hour and **+/D > 60** or **17** is their number. The required percentage of breakdowns is thus

```
100 * (+/D > 60) % #D
```

which has the value **18.89**.

The number of breakdowns less than **10** minutes in length is **+/D < 10** or **24**. Similarly, the number of breakdowns less than **20** minutes is **+/D < 20** or **46**, less than **30** minutes is **+/D < 30** or **59**, less than **40** minutes is **+/D < 40** or **66**, etc. Thus the number of breakdowns between **0** and **9** is **24**, between **10** and **19** is **22**, between **20** and **29** is **13**, between **30** and **39** is **7**, etc.

The expression **E = 1** which is equal to

```
0 0 1 1 0 0 0 0 0 0 ...
```

indicates that the third, fourth, ... failures were due to the CPU. Therefore the total number of CPU failures is **+/E = 1** or **19**. The total number of failures of the Disk drive is **+/E = 2** or **13**. Similarly the total number of failures for the Graph plotter, Printer, Servicing, and the Tape reader are **9**, **29**, **2** and **18**, respectively. The total number of failures is **+/19 13 9 29 2 18** or **90**. The total number of failures for each type of equipment may be found more simply by the expression

```
+/  E =/ 1 2 3 4 5 6
```

which is the list **19 13 9 29 2 18**. The percentage of failures of each type is given by

```
5.1": 100* 19 13 9 29 2 18 % 90
```

which is equal to

```
21.1 14.4 10.0 32.2  2.2 20.0
```

.

The calculations in this section require the following **J** verbs and adverbs in addition to those introduced in the previous section:

=	· *Equal*
<.	· *Lesser Of*
>.	· *Larger Of*
$	· *Shape*
#	· *Copy*
{	· *From*
":	· *Format*
/	· ***Table***
"	**Rank** ·

They may mentioned briefly during the presentation of the statistical material and illustrated at the end of the section with further examples and exercises.

Data summaries

Whereas all of the calculations in the previous section were done without the aid of defined verbs (except **am** for the arithmetic mean), we shall now use a number of defined verbs to continue our analysis of the computer breakdown data. These verbs are listed at the end of the section but their definition will not be discussed during their use. The interested reader may wish to study a few of them, possibly consulting Smillie (1995, 1966) where they are discussed in detail.

The downtimes for the Graph plotter are given by

```
G=. (E = 3) # X
```

which has the value

```
38 29 4 42 18 27 20 9 33
```

.

These times may be sorted by the expression **sort G** to give

```
4 9 18 20 27 29 33 38 42
```

.

A stem-and-leaf diagram is given by **SLitems G** and is

0	4 9
10	8
20	0 7 9
30	3 8
40	2

,

the interpretation of which is apparent from the above list of sorted frequencies.

The arithmetic mean of the downtimes is **am G** or **24.4**. An appreciation of the distribution of the data is given by what Sprent calls a five-statistic summary consisting of the minimum value, the first quartile, the second quartile or median, the third quartile and the maximum vlaue which for the present data are given by **five G** which is the five-item list

```
4 13.5 27 35.5 42
```

.

The stem-and-leaf diagram shown gives only those stem values corresponding to values greater than zero. If, for example, there were no downtimes between twenty and thirty-nine minutes, inclusive, the diagram would have only three rows corresponding to stem values of **0**, **10** and **40** minutes, and might give a misleading impression of the range of data. Therefore, it would be useful to be able to display the values in intervals over the complete range of the data. This may be accomplished by the verb **cfrtab** which gives a two-column frequency table with the midpoints of the intervals in the first column and the corresponding frequencies in the second. The expression

```
_0.5 10 5 cfrtab G
```

gives the following frequency table for the durations of the graph downtimes where there are **5** intervals of width **10** starting at **_0.5**:

```
 4.5 2
14.5 1
24.5 3
34.5 2
44.5 1
```

A frequency table for all of the downtimes **D** is given by

```
_0.5 10 4 cfrtab D
```

the first seven and last three rows of which are

```
 4.5 24
14.5 22
24.5 13
34.5  7
44.5  6
54.5  1
64.5  0
  .....
314.5  0
324.5  0
334.5  1
```

.

The following is a list of the defined verbs used in this section:

```
sort=. /:~
stem=. 10&* @ div&10
leaf=. 10&|
SLitems=. (~. @ stem  ;"0 stem </.
                leaf) @ sort
am=. +/ % #
q2=. [: am (<.,>.) @ -: @ <: @ # { ]
median=. q2 @ sort
q1=. q2 @ ((q2 > ]) # ])
Q1=. q1 @ sort
q3=. q2 @ ((q2 < ]) # ])
Q3=. q3 @ sort
five=. ({.,q1,q2,q3,{:) @ sort
cfrpts=. -:@(1&{) + {. + 1&{ * i.@{:
cfr=. i.@{:@[ rfr [: <. (] - {.@[) %
                                1&{@[
cfrtab=. cfrpts@[ ,. cfr
rfr=. +/"1 @ (=/)
```

Graphics

Statistical calculations in **J** may be supplemented by a number of different graphics packages. One very popular one is GNUPLOT, an interactive plotting program available by anonymous ftp at several sites. Commercially available spreadsheets such as Microsoft Excel or integrated packages such as Microsoft Works offer excellent graphics capabilities. Communication between **J** and whatever package is used is provided by the Clipboard which may be accessed, for example, using the verbs **CLIPwrite** and **CLIPread** discussed in Smillie (1995). The graphics given in this paper have been produced using Microsoft Works.

The barchart on the next page gives the frequency distribution of downtimes found in the previous section, and is followed by a pie chart giving the percentage frequencies for the various downtimes.

A feeling for numbers

Sprent devotes a chapter to a few simple examples intended to give the reader some experience in performing meaningful arithmetic in situations where a hasty decision or carelessly done calculations might give incorrect or misleading results. In this section we shall discuss some of these examples using **J** as a simple calculator so that we may concentrate our attention on the meaning of the calculations.

In one example we are asked to decide whether it is more profitable to invest £100 for one year at 10.5 per cent per annum compounded annually or at 10.3 per cent compounded semiannually. In the first scheme the **100** pounds will have increased to

```
100 + 0.105 * 100
```

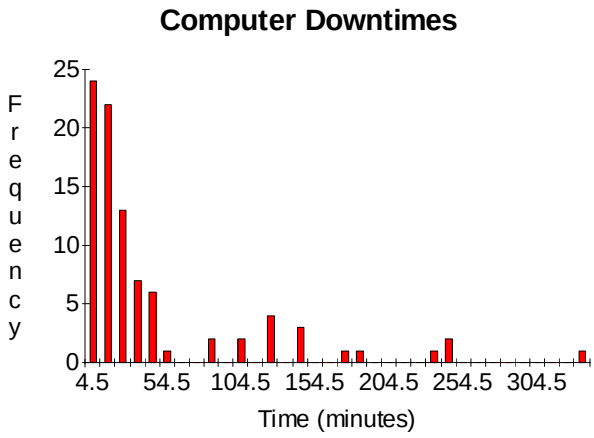
or **110.50** pounds. In the second scheme the **100** pounds will have increased to

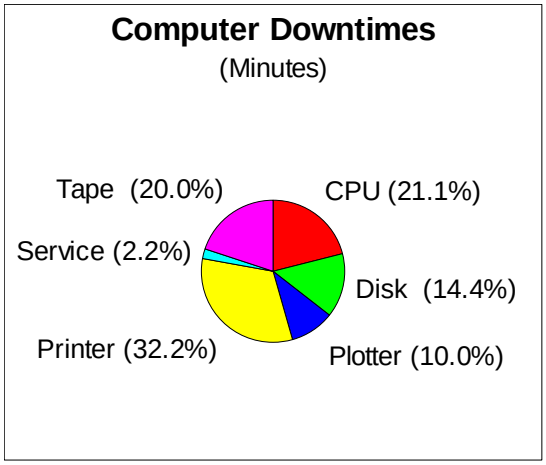
```
100 + 0.103 * 0.5 * 100
```

or **105.15** pounds at the end of six months, and this amount will have increased to

```
105.15 + 0.103 * 0.5 * 105.15
```

or **110.565** pounds at the end of the second six months. Therefore the second scheme is the more profitable by **110.57 - 110.50** or **7** pence.





It is of interest to find the annualized percentage rate equivalent to a given percentage rate componded monthly. For example, one pound owed at a monthly rate of 2 per cent becomes 1.02 pounds at the end of the first month, 1.02^2 at the end of 2 months, and eventually 1.02^{12} at the end of 12 months. Thus **1** pound at the end of a year has increased to **1.02 ^ 12** or **1.268** pounds, and the annualized interest rate equivalent to a monthly rate of **2** per cent is **26.8** per cent.

We shall calculate a short table of annualized percentage rates for a given range of monthly percentage rates given by the list **r** whose value is

```
1 1.5 2 2.5 3 4 .
```

The amount of one pound at the end of a month for each of these rates is

```
m1=. 1 + r % 100
```

or

```
1.01 1.015 1.02 1.025 1.03 1.04 ,
```

and the amounts after **12** months are

```
m12=. m1^12
```

or

```
1.12683 1.19562 1.26824 1.34489
          1.42576 1.60103  .
```

Therefore, the annualized percentage rates are

```
APR=. 100 * 1 | m12
```

which may be displayed in a two-column table together with the corresponding monthly rates by the expression

```
8.1 ": r ,. APR
```

which has the value

```
1.0    12.7
1.5    19.6
2.0    26.8
2.5    34.5
3.0    42.6
4.0    60.1  .
```

In order to give an idea of the effect of the size of numbers occurring during a calculation on the accuracy of the final result we are asked to consider

$$450 \times 449 \times 448 \times ... \times 440$$

divided by

$$461 \times 460 \times 459 \times ... \times 451$$

which is equal to 0.7644. Because of the size of the products that are accumulated while calculating each of the numerator and denominator, Sprent suggests that a reasonable order for performing the calculations would be 450 divided by 461, then multiplied by 449, then divided by 460, and continuing in this manner taking successive arguments alternately from the numerator and the denominator. The calculation may be done simply in **J** by the expression

```
(*/450 - i. 11) % */ 461 - i. 11
```

giving the value **0.7644359** which agrees with the correct value in Sprent.

The example in the last paragraph might raise the question of the number of significant digits retained in a calculation in **J**. One answer is given by examining a table of powers of **2** given by the expression

```
5.0 20.0 ": (] ,. 2&^) i. 56  .
```

The first five and last five rows of this table are

```
 0                    1
 1                    2
 2                    4
 3                    8
 4                   16
.....
51    2251799813685248
52    4503599627370496
53    9007199254740992
54   18014398509481980
55   36028797018963970
```

It is apparent that **2^54** cannot possibly be correct since the least significant digit of any positive power of two must be **2**, **4**, **6** or **8**. Furthermore the correctness of the powers of two up to and including **2^53** may be verified by consulting a book of mathematical tables or by manual calculation. Thus we see that calculations are performed

in **J** to an accuracy of about fifteen significant digits.

Simulation

The following table gives the number of typos found on each page of a nine-page manuscript:

```
1 2 3 4 5 6  7 8 9
0 3 1 9 4 7 11 8 2
```

The mean number of errors per page may be found simply as **(+/{:T151)% 9**, where **T151** represents the above table, and is equal to **5**. However, Sprent uses these data to introduce the concept of a sampling distribution and estimates the mean by finding the distribution of the mean number of errors in all **126** samples of size **4** pages taken from the **9** pages.

Let us introduce the defined verb **COMB** for giving all combinations of the items in the list right argument of length specified by the integer left argument, where, for example, **2 COMB 'cat'** is the table

```
ca
ct
at
```

.

All samples of four pages are generated very simply by the expression

```
c=. 4 COMB {:T151
```

which gives a table with **126** rows and **4** columns the first five rows of which are

```
0 3 1  9
0 3 1  4
0 3 1  7
0 3 1 11
0 3 1  8
```

.

The first row of this table gives the errors on pages **1**, **2**, **3** and **4**, the second the errors on pages **1**, **2**, **3** and **5**, etc. The mean number of errors in each of these combinations is given by **m=. am"1 c**, which is equal to

```
3.25 2 2.75 3.75 3 ...
```

showing that the mean number of errors on pages **1**, **2**, **3** and **4** is **3.25**, on pages **1**, **2**, **3** and **5** is **2**, etc. The mean of these means is **am m** or **5**. Finally, a frequency table of the means is given by

```
t=. 0.25 0.5 18 cfrtab m
```

.

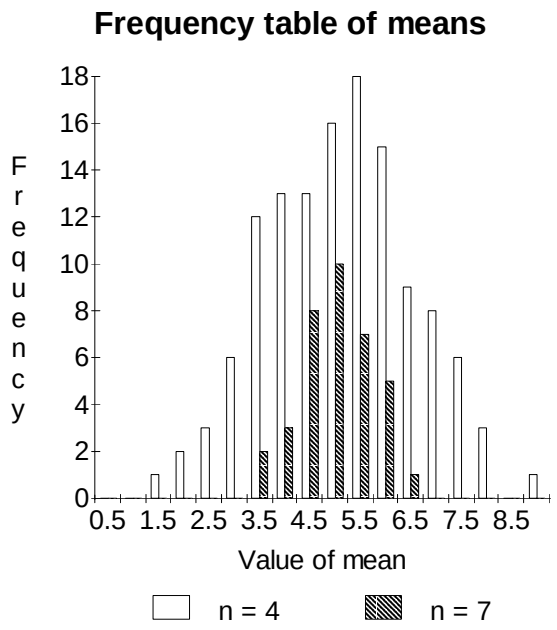
A similar sampling experiment but with the **36** samples of size **7** is given by the expression

```
t1=. 0.25 0.5 18 cfrtab m1=. am"1
                7 COMB {:T151
```

,

and we have that **am m1** is **5**.

These results may be used to give the following barchart:



It may be seen from this chart that there is much less variability in the means when they are estimated by samples of size **7** than by samples of size **4**.

The combinations verb **COMB** may be used without discussion in further sampling experiments. Also, if desired, its definition,

```
COMB=. ([ comb #@]) { ]
comb=. |.@ptt # i.@]
ptt=. (+/"1 @ tt@] e. [) # tt@]
tt=. #: @ i. @ (2&^)
```

,

may be used to introduce the important topics of permutations and combinations and of truth tables as given in Smillie (1995). An alternative, and more thorough, discussion of the first of these topics is given in Iverson (1991a).

References

Iverson, K. E., 1980. "The inductive method of introducing APL." *APL Users Meeting, Toronto, October 6, 7 8*. (Reprinted in *A Source Book of APL*, APL Press, Palo Alto, 1981.)

Iverson, K. E., 1991a. Arithmetic. Iverson Software Inc., Toronto.

Iverson, K. E., 1991b. "A personal view of APL." *IBM Systems Journal*, vol. 30, no. 4, pp. 582 - 593.

Iverson, K. E., 1994. ***J** Introduction and Dictionary*. Iverson Software Inc., Toronto.

Jenkins, Michael and Lynne Strugnell, 1992. *Teach Yourself Business Japanesse*. NTC Publishing Group, Lincolnwood, Ill.

Smillie, Keith, 1995. *Introducing J with Some Statistical Examples*. Available as a Postscript file by anonymous ftp at `ftp.cs.ualberta.ca` in the file `pub/smillie/intj.ps`.

Smillie, Keith, 1996. *Understanding Data with **J***. Available as a Postscript file by anonymous ftp at `ftp.cs.ualberta.ca` in the file `pub/smillie/undj.ps`.

Sprent, Peter, 1988. *Understanding Data*. Penguin Books, Ltd., Harmondsworth, Middlesex.

Appendix. Computer breakdowns for May 1984

1	842	4	49	13	1516	4	42	23	931	6	12
2	1035	2	18	14	1759	2	24	23	952	6	11
2	1529	1	123	15	803	4	7	23	1005	6	3
2	1735	1	3	15	817	4	9	23	1017	6	5
3	0	0	0	15	831	4	14	23	1024	6	9
4	1109	6	87	15	847	4	10	23	1038	6	3
4	1907	4	21	15	905	4	30	23	1402	1	22
4	2033	4	7	15	942	4	23	23	1823	1	17
4	2051	4	24	15	1009	4	141	24	0	0	0
5	835	4	11	15	1240	4	7	25	2129	3	20
5	922	4	19	15	1249	4	102	26	1340	6	5
6	900	5	243	15	1947	1	83	26	1545	6	3
7	1525	3	38	16	1429	3	29	26	1550	6	18
8	947	1	11	17	0	0	0	27	955	1	142
8	1010	1	18	18	0	0	0	27	1235	1	51
9	1432	2	32	19	1033	2	11	27	1707	2	18
9	1527	1	101	19	1102	2	15	27	1729	2	41
9	1712	1	6	19	1221	2	7	27	1944	4	9
9	1809	6	32	19	1234	2	6	28	1137	3	9
9	1931	4	27	19	1247	2	8	28	1430	6	27
9	2122	1	41	19	1342	2	121	28	1504	6	14
10	805	4	3	19	1829	3	4	29	913	3	33
11	1331	4	25	20	1328	6	23	29	952	1	128
11	1447	4	19	20	1700	3	42	29	1244	1	34
11	1702	6	242	20	1748	3	18	29	1359	6	18
12	0	0	0	21	904	6	31	29	1505	1	232
13	1022	1	122	21	1227	2	25	29	1901	1	179
13	1403	4	15	21	1513	4	42	29	2206	1	143
13	1422	4	7	21	1907	4	6	30	801	1	181
13	1433	4	6	22	800	5	339	30	1212	4	6
13	1445	4	2	22	1926	2	19	31	2105	6	14
13	1451	4	18	22	2240	3	27				

The columns represent date, time of breakdown, type of breakdown and length in minutes. The type of equipment is coded as follows: **0** No equipment, **1** CPU, **2** Disk drive, **3** Graph plotter, **4** Printer, **5** Service and **6** Tape reader.