MULTIVARIATE STATISTICS

A PRACTICAL APPROACH

Bernhard Flury
and
Hans Riedwyl

Multivariate Statistics
A practical approach

Bernhard Flury
and
Hans Riedwyl

Preface

During the last twenty years multivariate statistical methods have become increasingly popular among scientists in various fields. The theory had already made great progress in previous decades and routine applications of multivariate methods followed with the advent of fast computers. Nowadays statistical software packages perform in seconds what used to take weeks of tedious calculations.

Although this is certainly a welcome development, we find, on the other hand, that many users of statistical packages are not too sure of what they are doing, and this is especially true for multivariate statistical methods. Many researchers have heard about such techniques and feel intuitively that multivariate methods could be useful for their own work, but they haven't mastered the usual mathematical prerequisites. This book tries to fill the gap by explaining – in words and graphs – some basic concepts and selected methods of multivariate statistical analysis.

Why another book? Are the existing books on applied multivariate statistics all obsolete? No, some of them are up to date and, indeed, quite good. However, we think that this introduction distinguishes itself from any existing text in three ways. First, we illustrate the basic concepts mainly with graphical tools, hence the large number of figures. Second, all techniques, with just one exception, are illustrated by the same numerical data. Being familiar with the data will help the reader to understand what insights multivariate methods can offer. Third, we have avoided mathematical notation as much as possible. While we are well aware of the fact that this is a risky business, since avoiding mathematical language inevitably implies some loss of accuracy, we feel that it is possible to understand the basic ideas of multivariate analysis without mastering matrix algebra. Of course, we do not wish to discourage anybody from learning multivariate statistics the mathematical way. Indeed, many ideas and concepts can be stated in fewer words using concise mathematical language, but in order to appreciate the advantages of a more abstract approach the reader needs to be familiar with matrix theory, which most non-mathematicians tend to think of as a difficult subject.

This book has grown out of its German predecessor *Angewandte*

multivariate Statistik, published in 1983 by Gustav Fischer Verlag, Stuttgart. It is not just a translation, however; many new ideas and didactic tools have been introduced, one chapter has been removed, and another one has been added. The chapter that was deleted was the one on factor analysis, which was criticized by some reviewers of the German text. Although it is probably impossible to write about factor analysis without provoking criticism, we felt that we were not sufficiently expert in this field, and that it was probably better to omit this topic altogether.

Some remarks are in order concerning Chapter 5, which deals with multiple linear regression. This chapter differs from the others in having a more compressed presentation. It was indeed conceived as a restatement of linear regression rather than a first introduction, and assumes that the reader is familiar at least with simple linear regression. We felt that it wasn't possible to omit this chapter, since a basic understanding of linear regression is extremely helpful for understanding the linear discriminant function and related concepts.

We do not claim that our book treats the various multivariate methods comprehensively – on the contrary, we purposely limited the presentation to those ideas and techniques that seemed most important to us, and tried to explain those as carefully as possible. Our selection of topics is obviously biased towards our own research areas, but this is what we feel most competent to write about.

Many chapters are followed by one or several additional examples and by a discussion. The questions stem mostly from the participants of several courses on applied multivariate analysis that we taught from 1980 to 1985. These questions gave us an opportunity to discuss additional important methods or ideas that didn't fit into the main text. The length of the discussions essentially reflects the number of times the respective chapter has been taught so far.

Some readers of the German predecessor of this book have asked us to supplement it with detailed instructions for the use of existing statistical software such as BMDP, SAS, or SPSS. We have resisted the temptation to introduce such instructions for two reasons. First, we would like the information in this book to remain correct for, say, fifteen years. Most instructions for the use of currently existing programs will probably be obsolete by that time. Second, we think that statistical software should do what textbooks propose, whereas the current situation is often just the opposite: students think of statistics as the output produced by some computer program.

Understanding this book requires relatively few prerequisites: the reader should master basic algebra and be familiar with the basic methods of univariate statistics. Knowing linear regression, at least for the case of a single regressor, will be very helpful, however. It may be a good idea to do

some parallel reading in one of the books on regression given at the end of Chapter 5.

Among the many people who contributed to this book we name just a few. Rudolf Maibach wrote a preliminary version of Chapter 1. Erika Gautschi translated parts of the German text into English. George McCabe, Jean-Pierre Airoldi and three anonymous referees provided helpful comments on several chapters of the manuscript. Käthi Schütz and Emilia Bonnemain typed the manuscript with great enthusiasm, and Thomas Hänni programmed many graphics. We thank them all warmly. Finally, we would like to thank the publishers of this book, especially Elizabeth Johnston, and her colleagues at Chapman and Hall for their encouragement and cooperation.

Bernhard Flury
Hans Riedwyl
Berne, January 1987

Contents

1

The data

For the sake of simplicity we shall almost always refer to the same set of data for illustration. This comes from an inquiry that was conducted into genuine and forged thousand franc bills, henceforth called bills. How did we get hold of such unusual data?

To find the answer to this and other questions, let us listen for a while to two gentlemen sitting comfortably in their armchairs and chatting to each other. One is a statistician (S), the other an expert (E) in the fight against counterfeiting.

S: Now I'll tell you the real reason I invited you to my home this evening.

E: I thought the bottle of 1967 Gevrey-Chambertin was the reason, or am I wrong?

S: That was one reason, but not the only one. I would like to talk to you about an idea which I have been carrying around in my head for some time now.

E: Well, then – go ahead!

S: You have a lot of experience in sorting out genuine bills from forged ones, haven't you? You can tell, at first sight, whether a particular bill is genuine or not.

E: It's not quite that simple. I can certainly detect a *bad* forgery right away – say, if the water mark is missing, the paper is different, or the print contains gross errors. In sophisticated forgeries, however, one often has to look very closely to discover some small error. Forgers are becoming smarter all the time and their technical tools more and more precise and fine.

S: And so you have to rely on modern science to search for criminals.

E: Sure; but what are you driving at?

S: It occurred to me that statistics could help with this. In psychology and biology, for example, statistical methods are used for classifying items into groups and for determining which features characterize group membership.

E: That sounds very interesting, indeed. Can you be a bit more specific?

S: Alright, alright! I once examined a bill rather closely. What interests me are attributes that I can easily measure. The quality of the paper, the water mark, the colours are all attributes which do not satisfy these conditions. Linear dimensions, on the other hand, can easily be measured, even by a layman like myself. The forger, I am sure, has a lot of trouble reproducing all linear measures with perfect precision. So I could select various distances on the bill and measure them. For example, the length and height of the print image would be a good place to start.

E: Yes, sure, but how is this information going to help you?

S: I need it in order to compare genuine bills with forged ones. I could carry out the measurements on, let's say, 100 genuine and 100 forged bills. Since the forger surely does not work with absolute precision, I would find out very quickly that the measurements on the genuine bills differed from those on the forged ones – and that's precisely what I want to exploit.

E: Wait a minute! Why would you want to measure 100 bills; wouldn't a single one be enough?

S: No, because I am interested not only in the actual length of a line, but also in the variability of the measurements on a collection of bills. The variability between bills may give me information about the precision of the production process.

E: I see. One other question: is it at all possible to measure lines accurately enough in order to be able to detect a difference between genuine and forged bills?

S: That's a problem that can be solved. Either I can use sophisticated measurement equipment, or else I can try to project the bills on a screen and 'blow them up'. Naturally, there will still be some error of measurement, but it will be small compared with the variability due to the production process.

E: Alright, let's assume you have made all your measurements. Then you compare the two mean values for each given line. Correct?

S: It's not quite that simple. I can now carry out many different analyses. I consider the lengths that I have measured to be quantities subject to chance, or 'random variables'. I would like to have several of these, because the more random variables are available, the more information can be derived from them. To begin with, I would compare the genuine and the forged bills with regard to mean and standard deviation of each variable, and also with regard to correlations between the variables. Then I would represent, graphically, the various measurements of each bill. Furthermore, I'd like to know whether the group of forged bills differs markedly from the group of the genuine bills in the features under consideration. In other words, can we discriminate between the two groups? If so, I might be able to

tell with high probability whether a given bill is genuine or forged, just by combining the various measurements in an optimal way. In this way I could certainly provide you with valuable help in putting the forgers out of business.

E: All these methods you describe sound very good. But I'll tell you now how we approach the problem. First of all, you have to realize that I am not the one who encounters a forgery first. Only in the rarest of cases is it the police who discover the existence of a forgery. Generally it is the bank tellers, or other people who handle a lot of paper money, who first become suspicious of a bill. If the forgery is clumsy, they will detect it right away. If, however, the bill looks just a little bit unusual, they may send it for further examination. So, when it finally comes to me, it has already been established that a forgery is at hand. My first task consists of informing all banks and other money agencies about the forgery as quickly as possible. For this purpose a leaflet with a rough description of the falsification is sent to all interested parties. After that, we give one or several samples of the forgery to the printer who produces our bills and their specialists carry out a detailed examination. They check the bill with regard to printing process, type of paper, water mark, colours and chemical composition of inks, and much more. We expect this kind of information to help us find and eliminate the forgery workshop and the distribution organization.

S: Well, it is precisely in this investigation that I want to help you. On the basis of the features that I consider, I would like to determine whether a single bill can be attributed to a group or not. I call this 'identification'. Indeed, even during the production of genuine bills, one could, with the help of 'identification analysis', discover at an early stage a change in the manufacturing process, which would cause defective bills to be produced.

E: You probably assume that the forger manufactures the false notes under exactly the same homogeneous conditions as we manufacture our genuine bills. From experience I know that most of them are produced in batches in some old cellar or in a garage. For this reason I have my doubts about comparing forged and genuine bills statistically.

S: I, too, will not necessarily assume conditions of homogeneity. I will ask whether a collection of bills, for example a set of 100 forged ones, can be divided into classes in such a way that they look homogeneous within classes, but differ from one class to another. If that is the case, then of course I would like again to describe group differences. I hope I am not confusing you too much with these ideas.

E: Let's put this to a test! In the cause of science, I can certainly get you a few genuine and forged bills.
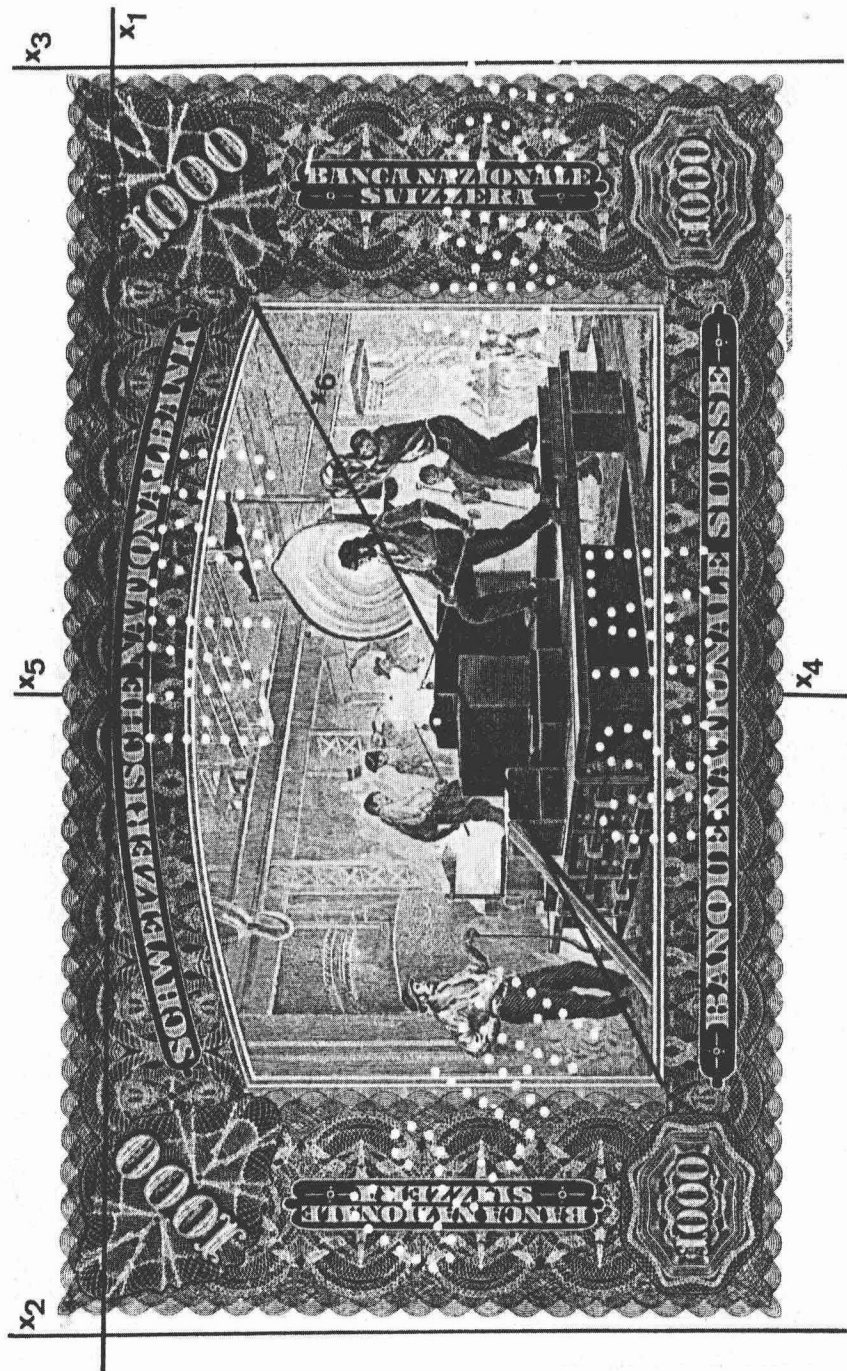
Figure 1.1 An old Swiss 1000-franc bill

Table 1.1 Six variables measured on 100 genuine Swiss bank notes

i	LENGTH x_{1i}	LEFT x_{2i}	RIGHT x_{3i}	BOTTOM x_{4i}	TOP x_{5i}	DIAGONAL x_{6i}
1	214.8	131.0	131.1	9.0	9.7	141.0
2	214.6	129.7	129.7	8.1	9.5	141.7
3	214.8	129.7	129.7	8.7	9.6	142.2
4	214.8	129.7	129.6	7.5	10.4	142.0
5	215.0	129.6	129.7	10.4	7.7	141.8
6	215.7	130.8	130.5	9.0	10.1	141.4
7	215.5	129.5	129.7	7.9	9.6	141.6
8	214.5	129.6	129.2	7.2	10.7	141.7
9	214.9	129.4	129.7	8.2	11.0	141.9
10	215.2	130.4	130.3	9.2	10.0	140.7
11	215.3	130.4	130.3	7.9	11.7	141.8
12	215.1	129.5	129.6	7.7	10.5	142.2
13	215.2	130.8	129.6	7.9	10.8	141.4
14	214.7	129.7	129.7	7.7	10.9	141.7
15	215.1	129.9	129.7	7.7	10.8	141.8
16	214.5	129.8	129.8	9.3	8.5	141.6
17	214.6	129.9	130.1	8.2	9.8	141.7
18	215.0	129.9	129.7	9.0	9.0	141.9
19	215.2	129.6	129.6	7.4	11.5	141.5
20	214.7	130.2	129.9	8.6	10.0	141.9
21	215.0	129.9	129.3	8.4	10.0	141.4
22	215.6	130.5	130.0	8.1	10.3	141.6
23	215.3	130.6	130.0	8.4	10.8	141.5
24	215.7	130.2	130.0	8.7	10.0	141.6
25	215.1	129.7	129.9	7.4	10.8	141.1
26	215.3	130.4	130.4	8.0	11.0	142.3
27	215.5	130.2	130.1	8.9	9.8	142.4
28	215.1	130.3	130.3	9.8	9.5	141.9
29	215.1	130.0	130.0	7.4	10.5	141.8
30	214.8	129.7	129.3	8.3	9.0	142.0
31	215.2	130.1	129.8	7.9	10.7	141.8
32	214.8	129.7	129.7	8.6	9.1	142.3
33	215.0	130.0	129.6	7.7	10.5	140.7
34	215.6	130.4	130.1	8.4	10.3	141.0
35	215.9	130.4	130.0	8.9	10.6	141.4
36	214.6	130.2	130.2	9.4	9.7	141.8
37	215.5	130.3	130.0	8.4	9.7	141.8
38	215.3	129.9	129.4	7.9	10.0	142.0
39	215.3	130.3	130.1	8.5	9.3	142.1
40	213.9	130.3	129.0	8.1	9.7	141.3
41	214.4	129.8	129.2	8.9	9.4	142.3
42	214.8	130.1	129.6	8.8	9.9	140.9
43	214.9	129.6	129.4	9.3	9.0	141.7
44	214.9	130.4	129.7	9.0	9.8	140.9
45	214.8	129.4	129.1	8.2	10.2	141.0
46	214.3	129.5	129.4	8.3	10.2	141.8
47	214.8	129.9	129.7	8.3	10.2	141.5
48	214.8	129.9	129.7	7.3	10.9	142.0
49	214.6	129.7	129.8	7.9	10.3	141.1
50	214.5	129.0	129.6	7.8	9.8	142.0
51	214.6	129.8	129.4	7.2	10.0	141.3
52	215.3	130.6	130.0	9.5	9.7	141.1
53	214.5	130.1	130.0	7.8	10.9	140.9
54	215.4	130.2	130.2	7.6	10.9	141.6
55	214.5	129.4	129.5	7.9	10.0	141.4
56	215.2	129.7	129.4	9.2	9.4	142.0

Table 1.1 (Continued)

i	LENGTH x_{1i}	LEFT x_{2i}	RIGHT x_{3i}	BOTTOM x_{4i}	TOP x_{5i}	DIAGONAL x_{6i}
57	215.7	130.0	129.4	9.2	10.4	141.2
58	215.0	129.6	129.4	8.8	9.0	141.1
59	215.1	130.1	129.9	7.9	11.0	141.3
60	215.1	130.0	129.8	8.2	10.3	141.4
61	215.1	129.6	129.3	8.3	9.9	141.6
62	215.3	129.7	129.4	7.5	10.5	141.5
63	215.4	129.8	129.4	8.0	10.6	141.5
64	214.5	130.0	129.5	8.0	10.8	141.4
65	215.0	130.0	129.8	8.6	10.6	141.5
66	215.2	130.6	130.0	8.8	10.6	140.8
67	214.6	129.5	129.2	7.7	10.3	141.3
68	214.8	129.7	129.3	9.1	9.5	141.5
69	215.1	129.6	129.8	8.6	9.8	141.8
70	214.9	130.2	130.2	8.0	11.2	139.6
71	213.8	129.8	129.5	8.4	11.1	140.9
72	215.2	129.9	129.5	8.2	10.3	141.4
73	215.0	129.6	130.2	8.7	10.0	141.2
74	214.4	129.9	129.6	7.5	10.5	141.8
75	215.2	129.9	129.7	7.2	10.6	142.1
76	214.1	129.6	129.3	7.6	10.7	141.7
77	214.9	129.9	130.1	8.8	10.0	141.2
78	214.6	129.8	129.4	7.4	10.6	141.0
79	215.2	130.5	129.8	7.9	10.9	140.9
80	214.6	129.9	129.4	7.9	10.0	141.8
81	215.1	129.7	129.7	8.6	10.3	140.6
82	214.9	129.8	129.6	7.5	10.3	141.0
83	215.2	129.7	129.1	9.0	9.7	141.9
84	215.2	130.1	129.9	7.9	10.8	141.3
85	215.4	130.7	130.2	9.0	11.1	141.2
86	215.1	129.9	129.6	8.9	10.2	141.5
87	215.2	129.9	129.7	8.7	9.5	141.6
88	215.0	129.6	129.2	8.4	10.2	142.1
89	214.9	130.3	129.9	7.4	11.2	141.5
90	215.0	129.9	129.7	8.0	10.5	142.0
91	214.7	129.7	129.3	8.6	9.6	141.6
92	215.4	130.0	129.9	8.5	9.7	141.4
93	214.9	129.4	129.5	8.2	9.9	141.5
94	214.5	129.5	129.3	7.4	10.7	141.5
95	214.7	129.6	129.5	8.3	10.0	142.0
96	215.6	129.9	129.9	9.0	9.5	141.7
97	215.0	130.4	130.3	9.1	10.2	141.1
98	214.4	129.7	129.5	8.0	10.3	141.2
99	215.1	130.0	129.8	9.1	10.2	141.5
100	214.7	130.0	129.4	7.8	10.0	141.2

S: That would be wonderful! I am sure that when I have completed my analysis and can show you the results you will find it much clearer. (Besides, many readers may then, hopefully, find pleasure in multivariate statistics.)

E: I hope you won't strain my modest mathematical and statistical knowledge too much.

Table 1.2 Six variables measured on 100 forged bank notes

i	LENGTH x_{1i}	LEFT x_{2i}	RIGHT x_{3i}	BOTTOM x_{4i}	TOP x_{5i}	DIAGONAL x_{6i}
101	214.4	130.1	130.3	9.7	11.7	139.8
102	214.9	130.5	130.2	11.0	11.5	139.5
103	214.9	130.3	130.1	8.7	11.7	140.2
104	215.0	130.4	130.6	9.9	10.9	140.3
105	214.7	130.2	130.3	11.8	10.9	139.7
106	215.0	130.2	130.2	10.6	10.7	139.9
107	215.3	130.3	130.1	9.3	12.1	140.2
108	214.8	130.1	130.4	9.8	11.5	139.9
109	215.0	130.2	129.9	10.0	11.9	139.4
110	215.2	130.6	130.8	10.4	11.2	140.3
111	215.2	130.4	130.3	8.0	11.5	139.2
112	215.1	130.5	130.3	10.6	11.5	140.1
113	215.4	130.7	131.1	9.7	11.8	140.6
114	214.9	130.4	129.9	11.4	11.0	139.9
115	215.1	130.3	130.0	10.6	10.8	139.7
116	215.5	130.4	130.0	8.2	11.2	139.2
117	214.7	130.6	130.1	11.8	10.5	139.8
118	214.7	130.4	130.1	12.1	10.4	139.9
119	214.8	130.5	130.2	11.0	11.0	140.0
120	214.4	130.2	129.9	10.1	12.0	139.2
121	214.8	130.3	130.4	10.1	12.1	139.6
122	215.1	130.6	130.3	12.3	10.2	139.6
123	215.3	130.8	131.1	11.6	10.6	140.2
124	215.1	130.7	130.4	10.5	11.2	139.7
125	214.7	130.5	130.5	9.9	10.3	140.1
126	214.9	130.0	130.3	10.2	11.4	139.6
127	215.0	130.4	130.4	9.4	11.6	140.2
128	215.5	130.7	130.3	10.2	11.8	140.0
129	215.1	130.2	130.2	10.1	11.3	140.3
130	214.5	130.2	130.6	9.8	12.1	139.9
131	214.3	130.2	130.0	10.7	10.5	139.8
132	214.5	130.2	129.8	12.3	11.2	139.2
133	214.9	130.5	130.2	10.6	11.5	139.9
134	214.6	130.2	130.4	10.5	11.8	139.7
135	214.2	130.0	130.2	11.0	11.2	139.5
136	214.8	130.1	130.1	11.9	11.1	139.5
137	214.6	129.8	130.2	10.7	11.1	139.4
138	214.9	130.7	130.3	9.3	11.2	138.3
139	214.6	130.4	130.4	11.3	10.8	139.8
140	214.5	130.5	130.2	11.8	10.2	139.6
141	214.8	130.2	130.3	10.0	11.9	139.3
142	214.7	130.0	129.4	10.2	11.0	139.2
143	214.6	130.2	130.4	11.2	10.7	139.9
144	215.0	130.5	130.4	10.6	11.1	139.9
145	214.5	129.8	129.8	11.4	10.0	139.3
146	214.9	130.6	130.4	11.9	10.5	139.8
147	215.0	130.5	130.4	11.4	10.7	139.9
148	215.3	130.6	130.3	9.3	11.3	138.1
149	214.7	130.2	130.1	10.7	11.0	139.4
150	214.9	129.9	130.0	9.9	12.3	139.4
151	214.9	130.3	129.9	11.9	10.6	139.8
152	214.6	129.9	129.7	11.9	10.1	139.0
153	214.6	129.7	129.3	10.4	11.0	139.3
154	214.5	130.1	130.1	12.1	10.3	139.4
155	214.5	130.3	130.0	11.0	11.5	139.5
156	215.1	130.0	130.3	11.6	10.5	139.7

Table 1.2 (Continued)

i	LENGTH x_{1i}	LEFT x_{2i}	RIGHT x_{3i}	BOTTOM x_{4i}	TOP x_{5i}	DIAGONAL x_{6i}
157	214.2	129.7	129.6	10.3	11.4	139.5
158	214.4	130.1	130.0	11.3	10.7	139.2
159	214.8	130.4	130.6	12.5	10.0	139.3
160	214.6	130.6	130.1	8.1	12.1	137.9
161	215.6	130.1	129.7	7.4	12.2	138.4
162	214.9	130.5	130.1	9.9	10.2	138.1
163	214.6	130.1	130.0	11.5	10.6	139.5
164	214.7	130.1	130.2	11.6	10.9	139.1
165	214.3	130.3	130.0	11.4	10.5	139.8
166	215.1	130.3	130.6	10.3	12.0	139.7
167	216.3	130.7	130.4	10.0	10.1	138.8
168	215.6	130.4	130.1	9.6	11.2	138.6
169	214.8	129.9	129.8	9.6	12.0	139.6
170	214.9	130.0	129.9	11.4	10.9	139.7
171	213.9	130.7	130.5	8.7	11.5	137.8
172	214.2	130.6	130.4	12.0	10.2	139.6
173	214.8	130.5	130.3	11.8	10.5	139.4
174	214.8	129.6	130.0	10.4	11.6	139.2
175	214.8	130.1	130.0	11.4	10.5	139.6
176	214.9	130.4	130.2	11.9	10.7	139.0
177	214.3	130.1	130.1	11.6	10.5	139.7
178	214.5	130.4	130.0	9.9	12.0	139.6
179	214.8	130.5	130.3	10.2	12.1	139.1
180	214.5	130.2	130.4	8.2	11.8	137.8
181	215.0	130.4	130.1	11.4	10.7	139.1
182	214.8	130.6	130.6	8.0	11.4	138.7
183	215.0	130.5	130.1	11.0	11.4	139.3
184	214.6	130.5	130.4	10.1	11.4	139.3
185	214.7	130.2	130.1	10.7	11.1	139.5
186	214.7	130.4	130.0	11.5	10.7	139.4
187	214.5	130.4	130.0	8.0	12.2	138.5
188	214.8	130.0	129.7	11.4	10.6	139.2
189	214.8	129.9	130.2	9.6	11.9	139.4
190	214.6	130.3	130.2	12.7	9.1	139.2
191	215.1	130.2	129.8	10.2	12.0	139.4
192	215.4	130.5	130.6	8.8	11.0	138.6
193	214.7	130.3	130.2	10.8	11.1	139.2
194	215.0	130.5	130.3	9.6	11.0	138.5
195	214.9	130.3	130.5	11.6	10.6	139.8
196	215.0	130.4	130.3	9.9	12.1	139.6
197	215.1	130.3	129.9	10.3	11.5	139.7
198	214.8	130.3	130.4	10.6	11.1	140.0
199	214.7	130.7	130.8	11.2	11.2	139.4
200	214.3	129.9	129.9	10.2	11.5	139.6

S: I will try, especially through graphical presentations, to explain things very simply.

And this is how we acquired the numerical data for our book. For didactic reasons we have chosen six linear measures on one side of the bill which capture roughly the size of the print image and paper format, as well as the position of the print image on the paper (see Fig. 1).