Multivariate Analysis of National Track Records

BRIAN DAWKINS*

A real data set of considerable intrinsic interest and offering considerable scope for investigation with different techniques of exploratory data analysis is examined using principal components analysis and the biplot. The analysis has an intuitively satisfying interpretation and illustrates well applications of the techniques. Plausible interpretations for the first and second principal components are suggested. A number of interesting aspects of the biplots are noted.

KEY WORDS: Biplot; Exploratory data analysis; Principal components analysis; Ranking.

1. INTRODUCTION

Recent analysis of track records has focused on either Olympic records (Chatterjee and Chatterjee 1982), on world records (Morton 1984), or physiological aspects of running (Lloyd 1983). Part of the problem in the analysis is the paucity of the data, particularly as relates to data for the performance of women. In fact, very little is made of the data for women in the cited references. Dyer (1982a,b, 1985) has gone some way toward redressing the balance.

A more extensive data base can be obtained by considering the national records for various track events (including the marathon). Indeed, from a data analytic viewpoint, this set offers a great deal to the analyst. Teachers of courses in multivariate analysis are perpetually on the lookout for good quality data sets that can be used to illustrate the various applied techniques used in exploratory analysis of high-dimension data sets, where high-dimensional may mean anything of dimension two or more. Of particular utility in this direction are sets for which the results of analysis are intuitively satisfying and indeed plausible. They serve to bolster one's faith that the techniques actually do something meaningful and worthwhile.

Such a data set can be constructed from the national records for men and women at various track races from 100 meters to the marathon, as given in Belcham and Hymans (1984). This is not a well-known data set, yet it can be analyzed instructively using a variety of techniques appropriate to exploratory data analysis (EDA), ranging from simple examination of the various one- and two-dimensional marginals, to more sophisticated methods, all of which lead to interesting results. I will consider here only the results of applying principal components analysis (PCA) (Chatfield and Collins 1980, chap. 4) and the biplot (Gabriel 1971, 1980, 1981; Everitt 1978). Such analyses are often an effective tool in the attempt to gain insight into high-dimensional data sets and, as noted by Chatfield and Collins (1980), make no probabilistic assumptions, concentrating on what are purely geometric properties of the data. Problems of inference are considered in Anderson (1984), but hardly seem appropriate in the present context.

*Brian Dawkins is Senior Lecturer, Mathematics Department, Victoria University of Wellington, Wellington, New Zealand.

2. THE ANALYSIS

For the purposes of the analysis, only a subset of the whole data set as reported in Belcham and Hymans (1984) was used. There are 55 countries for which a complete set of records for the "flat" races in both men's and women's events are available. Thus the hurdles and steeplechase events, as well as the field events, are excluded from the main analysis, although for purposes of comparison, more complete data sets are referenced at certain points. The full data set has many missing values, and handling these is an important and interesting problem, but for the purposes of this paper I will restrict attention to the countries with a complete set of observations. The men's and women's sets will be treated separately. Thus the data matrix for the women was 55×7, with the events represented being the 100 meters, 200 meters, 400 meters, 800 meters, 1,500 meters, 3,000 meters, and marathon. For the men the data matrix was 55×8, differing from the women's events in that the 3,000 meters was excluded but the 5,000 and 10,000 meters were included.

The first step in the analysis was to rescale each data set

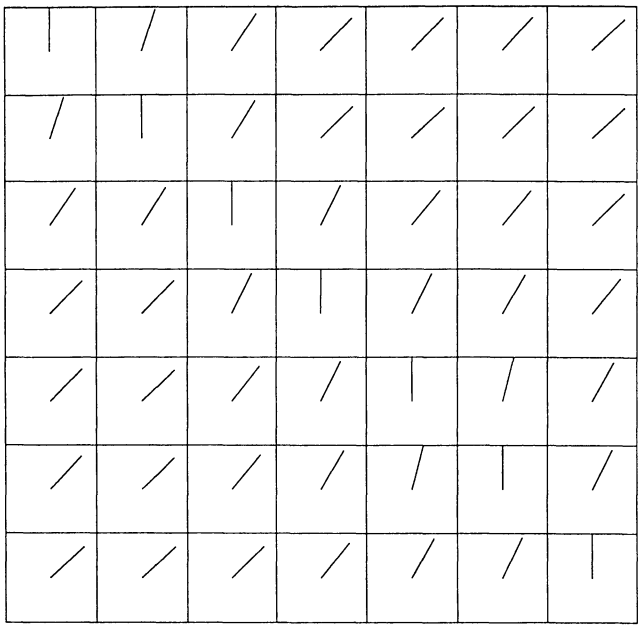

Figure 1. The Correlation Grid Representing the Women's Data. This grid is a graphical representation of the correlation matrix C of the women's data. Each cell in the grid corresponds in an obvious way to an element of C. The stub vectors in each cell are of the same length, and their directions represent the values of C in that the cosine of the angle with the vertical is the corresponding correlation. Thus the nearer the vertical, the higher the absolute value of the correlation, whereas horizontal vectors represent zero correlations. Patterns such as the monotone decrease of correlation between events as the distances become more disparate are easily seen.

Table 1. Rankings and Scores for Women Based on the First Principal Components

Track ranking	Nation	Track score	Olympic rank	Olympic score
1	German Democratic Republic	3.51	1	6.19
2	U.S.S.R.	3.47	2	5.82
3	U.S.A.	3.35	3	5.05
4	Czechoslovakia	3.06	5	4.74
5	Federal Republic of Germany	2.93	4	4.71
6	Great Britain and Northern Ireland	2.71	6	4.38
7	Poland	2.68	7	4.23
8	Canada	2.61	11	3.29
9	Finland	2.19	10	3.38
10	Italy	2.14	14	2.63

to give mean 0 and standard deviation 1, on the grounds that all of the variables are equally important. If the raw data were analyzed using the same time units, the marathon data would completely swamp the effect of the shorter races and would, therefore, be weighted excessively in the analysis.

One of the objectives of the analysis was to seek some sort of objective measure of the athletic excellence of a given nation with a view to ranking them, much as tennis players are ranked.

PCA was carried out on the data sets as detailed previously. For the women's data, the standard deviations for the principal components were 2.41, .81, .55, .35, .23, .20, and .15, with the first component accounting for 83% of the variation and the first two components accounting for 92%. Investigation of the first principal component suggested that it was just such a criterion as sought. It is essentially a normalized unit vector, and although this is to some extent a reflection of the normalizing procedure, the uniformity of the weights is really a reflection of the considerable amount of structure in the original data. Using it as a basis for ranks produced the ordering of the first 10 nations as shown in Table 1. The table also gives the scores on which the rankings were based, together with scores and ranks calculated from data on the complete set of records, including all Olympic events, both track and field, for women. This latter set forms a 50 × 15 array for which the first principal component explains 79% of the variation, the first two components explaining 86%. The first principal component is again essentially a multiple of the unit vector.

The first two columns of the rotation matrix based on track results only are given in Table 2. Note that here I am interpreting the matrix formed by the eigenvectors as the orthogonal matrix rotating the original data into the principal components orientation (Becker and Chambers 1984, p. 396; Chatfield and Collins 1980, p. 57). All principal component computations are the results of applying the S function prcomp to the appropriate data set. It is interesting to note that there is a misprint in Anderson (1984, p. 451) that identifies principal components incorrectly with the eigenvectors of the covariance matrix, instead of with the linear combinations based on those eigenvectors.

Table 2. First Two Columns of Rotation Matrices for the PCA of Both the Men's and the Women's Data

Women		Men	
First principal component	Second principal component	Second principal component	First principal component
.37	−.49	.32	−.57
.37	−.54	.34	−.46
.38	−.25	.36	−.25
.38	.15	.37	−.01
.40	.36	.37	.14
.39	.35	.36	.31
.37	.37	.37	.31
		.34	.44

NOTE: $1/\sqrt{7}$ is .38 to two decimal places and hence the first principal component in the case of the women is essentially a normalized unit vector. A corresponding observation can be made about the first principal component for the men's data, given that $1/\sqrt{8}$ = .35 to two decimal places.

The first two columns as given are of some considerable interest, with remarkably uniform weightings in the first principal component. The second principal component appears to be interpretable as a measure of relative strength of a given nation at the various distances. Hence a score based on the second component that is near 0 would seem to indicate that the particular nation had achieved at about the same level in both long and short distances, etcetera. This idea will be further elaborated after a brief discussion of the men's results.

Similar analysis of the men's data gave values for the

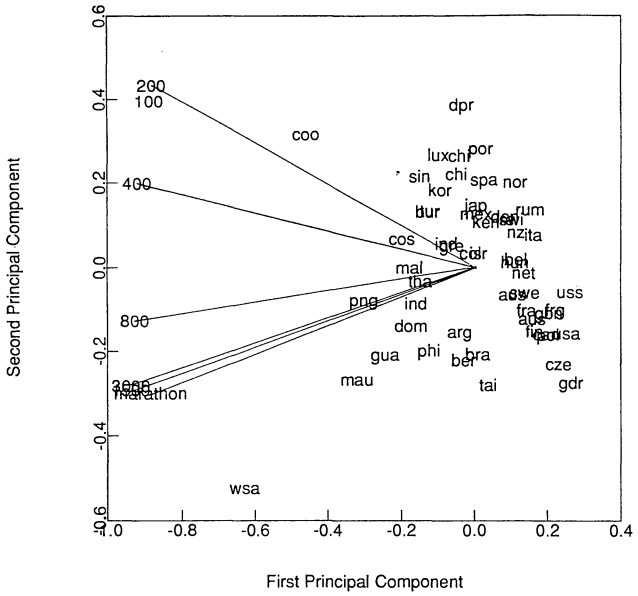

Figure 2. Biplot of the Data for Women. The three-letter codes denote the countries for which the data were available. Although this leads to considerable overprinting, the general picture is quite clear, and many countries are easily identifiable, even without a key. The close correlation of the longer events is clearly demonstrated by the tight cluster of vectors representing the corresponding variables. Overall rankings of the countries can be derived from the first principal component values. Further interpretations are given in the body of the article.

Table 3. Rankings and Scores for Men Based on the First Principal Components

Track ranking	Nation	Track score	Olympic rank	Olympic score	Olympic second component
1	U.S.A.	3.43	1	8.09	1.51
2	Great Britain and Northern Ireland	3.02	5	6.30	.15
3	Italy	2.73	7	5.67	.73
4	U.S.S.R.	2.63	2	6.80	.90
5	German Democratic Republic	2.59	3	6.72	.52
6	Federal Republic of Germany	2.55	4	6.57	.44
7	Australia	2.45	12	4.76	−.09
8	France	2.17	8	5.49	.57
9	Kenya	2.17	27	2.68	−1.08
10	Belgium	2.04	19	3.93	−.58

standard deviations of the principal components of 2.57, .94, .40, .35, .28, .26, .21, and .15, with the first principal component accounting for 82% of the variation. The first two principal components accounted for 94% of the variation.

The first two columns of the rotation matrix based on the scaled data are given in Table 2 and have much the same interpretation as those derived from the women's data. Rankings based on weights from the first principal component are as shown in Table 3.

Included in Table 3 are results from an analysis based on a more extensive data set that included those nations having a complete set of records in the Olympic events. This is a 90 × 22 data set and gave rankings that in most cases were consistent with those from the track records alone. The appropriate ranks and scores are given in Table 3, which also includes second component values.

3. DISCUSSION

The analysis by PCA of the data sets has resulted in rankings that are at least plausible from the viewpoint of subjective judgments, and, in any case, these have a considerable intrinsic interest. This is a considerable merit when one is attempting to teach such a technique. It also tends to support one's belief in the utility of the technique in areas where such judgments are not so easy to come by.

The rankings are essentially ordinary means of the suitably normalized data, and this reflects the specialized nature of the geometry of the data set. Roughly speaking, the data points are distributed about a line in the appropriately dimensioned space. Thus, in the track data, examination of the individual scatterplots reveals the existence of linear correlation. The correlation is strongest when the ratio of the longer to the shorter distances involved is least.

As is observed in Chatfield and Collins (1980, p. 72), it is well known that strong correlation between the variables implies that the first principal component can be regarded as a measure of size in some sense. Since the smallest correlation in the women's data is .69, it is not surprising to find quite strong relationships emerging among the variables. Figure 1 graphically displays the values of the correlation matrix for the data.

In this case, the second principal component is of considerable interest since it seems to have a fairly clear interpretation as a measure of differential achievement, in the sense that values around 0 seem to indicate achievement at about the same level internationally, whereas extreme values in either direction indicate an imbalance in achievement. Thus, for example, looking at Table 3, we see that although in the track rankings Kenya is rated as eighth equal, in the overall rankings based on the Olympic events it is only rated twenty-seventh, and this is reflected by the negative value of the second component. This is indicative of its distinguished record on the track but relatively poor achievement in the field events. The U.S.A., on the other hand, has a second component value of 1.51, reflecting its dominance in certain events.

4. THE BIPLOT

I will now change viewpoints slightly and look at some of the same data using the biplot as discussed in Everitt (1978). As noted there, the technique involves the factorization of any $m \times n$ matrix Y of rank r as $Y = GH'$, where G is $n \times r$ and H is $m \times r$, where both G and H are of rank r. If r is greater than 2, suitable rank 2 approximations to each of G and H are derived. These approximations have two independent columns, and the rows of these two matrices are taken as representing n and m points in the plane, respectively. The biplot is then the scatterplot of the points represented by the rows. One useful rank 2

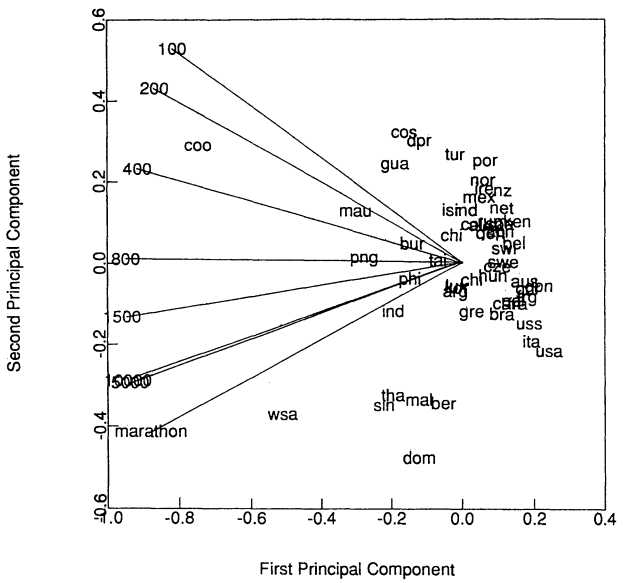

Figure 3. Biplot of the Data for Men. Close correlation of the 5,000 m and 10,000 m events is indicated by the superposition of the vectors corresponding to these variables. As in Figure 2, overall rankings may be derived from the first principal component values. Interestingly enough, comparison with Figure 2 reveals that there is rough comparability in the placing of the codes representing the various countries, as well as in the disposition of the vectors representing the events.

approximation is based on the eigenvalue analysis of the matrix of sums of squares and cross-products based on the data matrix, though generally it is necessary to introduce some form of normalization to get sensible biplots. This is particularly so if the original variables have disparate scales of measurement. This is the case with the track and field data discussed in this article, and all computations are based on appropriately centered and scaled data.

With a rank 2 approximation of the type given previously, some useful geometric interpretations can be placed on the scatterplots. An approximation of the original data matrix can be generated by forming the appropriate products of the rows of G with the rows H, and the cosines of the angles between the variable vectors can be generated as an approximation of the correlations between the original variables. See Everitt (1978) for other properties.

5. DISCUSSION

Thus, in Figure 2, the biplot based on the 55×7 data matrix of track data for the women, the decreasing correlation between the marathon and the other events as they decrease in distance is clearly visible. In Figure 2, the countries involved are indexed by a three-letter code, and although this involves some overwriting, the overall picture is fairly clear.

Figure 3 has a comparable construction based on the 55×8 matrix of men's data, and it is broadly similar in nature,

Table 4. National Records for Women

Country	100m (secs)	200m (secs)	400m (secs)	800m (mins)	1,500m (mins)	3,000m (mins)	Marathon (mins)
argentina	11.61	22.94	54.50	2.15	4.43	9.79	178.52
australia	11.20	22.35	51.08	1.98	4.13	9.08	152.37
austria	11.43	23.09	50.62	1.99	4.22	9.34	159.37
belgium	11.41	23.04	52.00	2.00	4.14	8.88	157.85
bermuda	11.46	23.05	53.30	2.16	4.58	9.81	169.98
brazil	11.31	23.17	52.80	2.10	4.49	9.77	168.75
burma	12.14	24.47	55.00	2.18	4.45	9.51	191.02
canada	11.00	22.25	50.06	2.00	4.06	8.81	149.45
chile	12.00	24.52	54.90	2.05	4.23	9.37	171.38
china	11.95	24.41	54.97	2.08	4.33	9.31	168.48
colombia	11.60	24.00	53.26	2.11	4.35	9.46	165.42
cookis	12.90	27.10	60.40	2.30	4.84	11.10	233.22
costa	11.96	24.60	58.25	2.21	4.68	10.43	171.80
czech	11.09	21.97	47.99	1.89	4.14	8.92	158.85
denmark	11.42	23.52	53.60	2.03	4.18	8.71	151.75
domrep	11.79	24.05	56.05	2.24	4.74	9.89	203.88
finland	11.13	22.39	50.14	2.03	4.10	8.92	154.23
france	11.15	22.59	51.73	2.00	4.14	8.98	155.27
gdr	10.81	21.71	48.16	1.93	3.96	8.75	157.68
frg	11.01	22.39	49.75	1.95	4.03	8.59	148.53
gbni	11.00	22.13	50.46	1.98	4.03	8.62	149.72
greece	11.79	24.08	54.93	2.07	4.35	9.87	182.20
guatemala	11.84	24.54	56.09	2.28	4.86	10.54	215.08
hungary	11.45	23.06	51.50	2.01	4.14	8.98	156.37
india	11.95	24.28	53.60	2.10	4.32	9.98	188.03
indonesia	11.85	24.24	55.34	2.22	4.61	10.02	201.28
ireland	11.43	23.51	53.24	2.05	4.11	8.89	149.38
israel	11.45	23.57	54.9	2.10	4.25	9.37	160.48
italy	11.29	23.00	52.01	1.96	3.98	8.63	151.82
japan	11.73	24.00	53.73	2.09	4.35	9.20	150.50
kenya	11.73	23.88	52.70	2.00	4.15	9.20	181.05
korea	11.96	24.49	55.70	2.15	4.42	9.62	164.65
dprkorea	12.25	25.78	51.20	1.97	4.25	9.35	179.17
luxembourg	12.03	24.96	56.10	2.07	4.38	9.64	174.68
malaysia	12.23	24.21	55.09	2.19	4.69	10.46	182.17
mauritius	11.76	25.08	58.10	2.27	4.79	10.90	261.13
mexico	11.89	23.62	53.76	2.04	4.25	9.59	158.53
netherlands	11.25	22.81	52.38	1.99	4.06	9.01	152.48
nz	11.55	23.13	51.60	2.02	4.18	8.76	145.48
norway	11.58	23.31	53.12	2.03	4.01	8.53	145.48
png	12.25	25.07	56.96	2.24	4.84	10.69	233.00
philippines	11.76	23.54	54.60	2.19	4.60	10.16	200.37
poland	11.13	22.21	49.29	1.95	3.99	8.97	160.82
portugal	11.81	24.22	54.30	2.09	4.16	8.84	151.20
rumania	11.44	23.46	51.20	1.92	3.96	8.53	165.45
singapore	12.30	25.00	55.08	2.12	4.52	9.94	182.77
spain	11.80	23.98	53.59	2.05	4.14	9.02	162.60
sweden	11.16	22.82	51.79	2.02	4.12	8.84	154.48
switzerland	11.45	23.31	53.11	2.02	4.07	8.77	153.42
taipei	11.22	22.62	52.50	2.10	4.38	9.63	177.87
thailand	11.75	24.46	55.80	2.20	4.72	10.28	168.45
turkey	11.98	24.44	56.45	2.15	4.37	9.38	201.08
usa	10.79	21.83	50.62	1.96	3.95	8.50	142.72
ussr	11.06	22.19	49.19	1.89	3.87	8.45	151.22
wsamoa	12.74	25.85	58.73	2.33	5.81	13.04	306.00

Source: IAAF/ATFS Track and Field Statistics Handbook for the 1984 Los Angeles Olympics.

including the relative positioning of the various countries involved. Again, the different correlations between the events can be estimated at a glance, and there are some interesting aspects of these. For example, although for the women there is very tight correlation between the 3,000 meters and the marathon, for the men there is a distinct difference between the marathon and the long track races. This probably reflects the fact that men tend to specialize, so a given athlete will run the 5,000 and 10,000 meters but not the marathon, or vice versa, whereas the women quite often run both shorter and longer distances.

Again, the dominance of U.S. male sprinters is clearly depicted by the relation of the point of the scatterplot denoting the U.S.A. to the vectors corresponding to the 100 meters and 200 meters. Similar interpretations can be placed on the relations of other point and variable vectors.

6. THE DATA SET

As noted previously, the data used, as given in Tables 4 and 5, form a very small part of the large data base available in Belcham and Hymans (1984). This volume is from the catalog of ongoing publications sponsored by the IAAF, a comprehensive list of whose publications can be obtained by writing to the Publications Officer, International Amateur Athletic Federation, 3 Hans Crescent, Knightsbridge, London, SW1X OLN, England. These publications are reasonably priced and, as indicated, have enormous amounts of interesting data, well worth examining.

Table 5. National Records for Men

Country	100m (secs)	200m (secs)	400m (secs)	800m (mins)	1,500m (mins)	5,000m (mins)	10,000m (mins)	Marathon (mins)
argentina	10.39	20.81	46.84	1.81	3.70	14.04	29.36	137.72
australia	10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.30
austria	10.44	20.81	46.82	1.79	3.60	13.26	27.72	135.90
belgium	10.34	20.68	45.04	1.73	3.60	13.22	27.45	129.95
bermuda	10.28	20.58	45.91	1.80	3.75	14.68	30.55	146.62
brazil	10.22	20.43	45.21	1.73	3.66	13.62	28.62	133.13
burma	10.64	21.52	48.30	1.80	3.85	14.45	30.28	139.95
canada	10.17	20.22	45.68	1.76	3.63	13.55	28.09	130.15
chile	10.34	20.8	46.20	1.79	3.71	13.61	29.30	134.03
china	10.51	21.04	47.30	1.81	3.73	13.90	29.13	133.53
colombia	10.43	21.05	46.10	1.82	3.74	13.49	27.88	131.35
cookis	12.18	23.2	52.94	2.02	4.24	16.70	35.38	164.70
costa	10.94	21.9	48.66	1.87	3.84	14.03	28.81	136.58
czech	10.35	20.65	45.64	1.76	3.58	13.42	28.19	134.32
denmark	10.56	20.52	45.89	1.78	3.61	13.50	28.11	130.78
domrep	10.14	20.65	46.80	1.82	3.82	14.91	31.45	154.12
finland	10.43	20.69	45.49	1.74	3.61	13.27	27.52	130.87
france	10.11	20.38	45.28	1.73	3.57	13.34	27.97	132.30
gdr	10.12	20.33	44.87	1.73	3.56	13.17	27.42	129.92
frg	10.16	20.37	44.50	1.73	3.53	13.21	27.61	132.23
gbni	10.11	20.21	44.93	1.70	3.51	13.01	27.51	129.13
greece	10.22	20.71	46.56	1.78	3.64	14.59	28.45	134.60
guatemala	10.98	21.82	48.40	1.89	3.80	14.16	30.11	139.33
hungary	10.26	20.62	46.02	1.77	3.62	13.49	28.44	132.58
india	10.60	21.42	45.73	1.76	3.73	13.77	28.81	131.98
indonesia	10.59	21.49	47.80	1.84	3.92	14.73	30.79	148.83
ireland	10.61	20.96	46.30	1.79	3.56	13.32	27.81	132.35
israel	10.71	21.00	47.80	1.77	3.72	13.66	28.93	137.55
italy	10.01	19.72	45.26	1.73	3.60	13.23	27.52	131.08
japan	10.34	20.81	45.86	1.79	3.64	13.41	27.72	128.63
kenya	10.46	20.66	44.92	1.73	3.55	13.10	27.38	129.75
korea	10.34	20.89	46.90	1.79	3.77	13.96	29.23	136.25
dprkorea	10.91	21.94	47.30	1.85	3.77	14.13	29.67	130.87
luxembourg	10.35	20.77	47.40	1.82	3.67	13.64	29.08	141.27
malaysia	10.40	20.92	46.30	1.82	3.80	14.64	31.01	154.10
mauritius	11.19	22.45	47.70	1.88	3.83	15.06	31.77	152.23
mexico	10.42	21.30	46.10	1.80	3.65	13.46	27.95	129.20
netherlands	10.52	20.95	45.10	1.74	3.62	13.36	27.61	129.02
nz	10.51	20.88	46.10	1.74	3.54	13.21	27.70	128.98
norway	10.55	21.16	46.71	1.76	3.62	13.34	27.69	131.48
png	10.96	21.78	47.90	1.90	4.01	14.72	31.36	148.22
philippines	10.78	21.64	46.24	1.81	3.83	14.74	30.64	145.27
poland	10.16	20.24	45.36	1.76	3.60	13.29	27.89	131.58
portugal	10.53	21.17	46.70	1.79	3.62	13.13	27.38	128.65
rumania	10.41	20.98	45.87	1.76	3.64	13.25	27.67	132.50
singapore	10.38	21.28	47.40	1.88	3.89	15.11	31.32	157.77
spain	10.42	20.77	45.98	1.76	3.55	13.31	27.73	131.57
sweden	10.25	20.61	45.63	1.77	3.61	13.29	27.94	130.63
switzerland	10.37	20.46	45.78	1.78	3.55	13.22	27.91	131.20
taipei	10.59	21.29	46.80	1.79	3.77	14.07	30.07	139.27
thailand	10.39	21.09	47.91	1.83	3.84	15.23	32.56	149.90
turkey	10.71	21.43	47.6	1.79	3.67	13.56	28.58	131.50
usa	9.93	19.75	43.86	1.73	3.53	13.20	27.43	128.22
ussr	10.07	20.00	44.6	1.75	3.59	13.20	27.53	130.55
wsamoa	10.82	21.86	49.00	2.02	4.24	16.28	34.71	161.83

Source: IAAF/ATFS Track and Field Statistics Handbook for the 1984 Los Angeles Olympics.

7. CONCLUSION

I hope that enough has been indicated to show how the data set is richly rewarding to analyze without being overly complicated in basic structure. It can be used to illustrate one- and two-dimensional marginals of varying types, as well as such things as canonical correlation and multiple regression, including collinearity and influential observations, clustering, Q–Q plots, and many more. It deserves on several grounds to be better known in the statistical education literature.

Such analyses as have been presented here are easily carried out in the sort of environment offered by S (Becker and Chambers 1984). All of the results and graphics in this article have been obtained using S running on a 3B2/400+.

[*Received June 1988. Revised September 1988.*]

REFERENCES

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis* (2nd ed.), New York: John Wiley.

Becker, R. A., and Chambers, J. M. (1984), *S: An Interactive Environment for Data Analysis and Graphics*, Belmont, CA: Wadsworth.

Belcham P., and Hymans, R. (eds.) (1984), *IAAF/ATFS Track and Field Statistics Handbook for the 1984 Los Angeles Olympic Games*, London: International Amateur Athletic Federation.

Chatfield, C., and Collins, A. J. (1980), *Introduction to Multivariate Analysis*, London: Chapman & Hall.

Chatterjee, S., and Chatterjee, S. (1982), "New Lamps for Old: An Exploratory Analysis of Running Times in the Olympic Games," *Journal of the Royal Statistical Society*, Ser. C, 31, 14–22.

Dyer, K. F. (1982a), *Catching Up With the Men*, London: Junction Books.

——— (1982b), *Challenging the Men: The Social Biology of Female Sporting Achievement*, St. Lucia, Queensland: University of Queensland Press.

——— (1985), "Making Up the Difference," *Search*, 16, No. 9-12, 264–269.

Everitt, B. S. (1978), *Graphical Techniques for Multivariate Data Analysis*, New York: North-Holland.

Gabriel, K. R. (1971), "The Biplot-Graphic Display of Matrices With Applications to Principal Components Analysis," *Biometrika*, 58, 453–467.

——— (1980), "Biplot," in *Encyclopaedia of Statistical Sciences* (Vol. 1), eds. N. L. Johnson and S. Kotz, New York: John Wiley, pp. 263–271.

——— (1981), "Biplot Display of Multivariate Matrices for Inspection of Data and Diagnosis," in *Interpreting Multivariate Data*, ed. V. Barnett, New York: John Wiley, pp. 147–174.

Lloyd, B. B. (1983), "Runs and Sums: The Application of Mathematics to the Analysis of Running Records," in *Maths at Work*, eds. A. G. Howson and R. McLone, London: Heinemann Educational Books, pp. 65–78.

Morton, R. H. (1984), Letter to the Editor, *Journal of the Royal Statistical Society*, Ser. C, 33, 317–318.

The Gamma Distribution as a Mixture of Exponential Distributions

LEON JAY GLESER*

A gamma distribution with arbitrary scale parameter θ and shape parameter $r < 1$ can be represented as a scale mixture of exponential distributions.

KEY WORDS: Decreasing hazard rate; Pooled data; Tests of fit.

1. INTRODUCTION AND MAIN RESULT

In a paper by Proschan (1963) on failure rate analysis some data are included concerning the time of successive failures of the air conditioning system of each member of a fleet of 13 Boeing 720 jet airplanes. Proschan tested the fit of the exponential distribution to these data using the Kolmogorov–Smirnov test of fit and was unable to reject the hypothesis that the pooled data are exponentially distributed. Proschan remarked, however, that the pooled data seemed to exhibit a decreasing failure rate, and he thus questioned whether the exponential distribution really does provide an adequate model for the data.

In a later paper, Dahiya and Gurland (1972) used a test based on the sample moments to test the fit of the exponential distribution in Proschan's data against gamma alternatives. Their test rejected the null hypothesis of exponentiality at the .01 level of significance, confirming Proschan's doubts. They found that a gamma distribution with scale parameter $\theta = (122.56)^{-1}$ and shape parameter $r = .76$ provides a good fit to Proschan's data. They noted that such a gamma distribution has a decreasing failure rate.

In Olkin, Gleser, and Derman (1980), Proschan's data were used as an example of data that appear to follow an exponential distribution. In preparing a revision of Olkin et al. (1980), I came across Dahiya and Gurland's paper and became interested in how I could explain their conclusions. Since Proschan had combined data from several airplanes, which might be subject to different uses and environments, it was natural to suspect (as Proschan had) that survival times might have different exponential distributions for different planes and thus that Proschan's data would follow a mixture of exponential distributions. This led to the question

*Leon Jay Gleser is Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907. Research for this article was supported by National Science Foundation Grant DMS-8501966.