Chapter 3

Preliminary Data Analysis

This chapter covers the preliminary data analysis. We study the winning results for Olympic sports champions over the period 1928-2000. We look at three sports (swimming for 299 results from 26 events, running for 188 results from 13 events, jumping for 84 results from six events throwing for 115 results from seven events). In each case the outcome is the winning time (or distance for jumping and throwing events). The determinants are gender, year, and type of event. The names and countries represented by the winners are also recorded in the database. Our aim is to develop a predictive model for the winning result in each event. We begin by examining the swimming events, and then extend our results to the other sports.

3.1 Description of Database

The data are stored in three database tables using Ms Access.
Figure 3.1 shows the structure of this database.

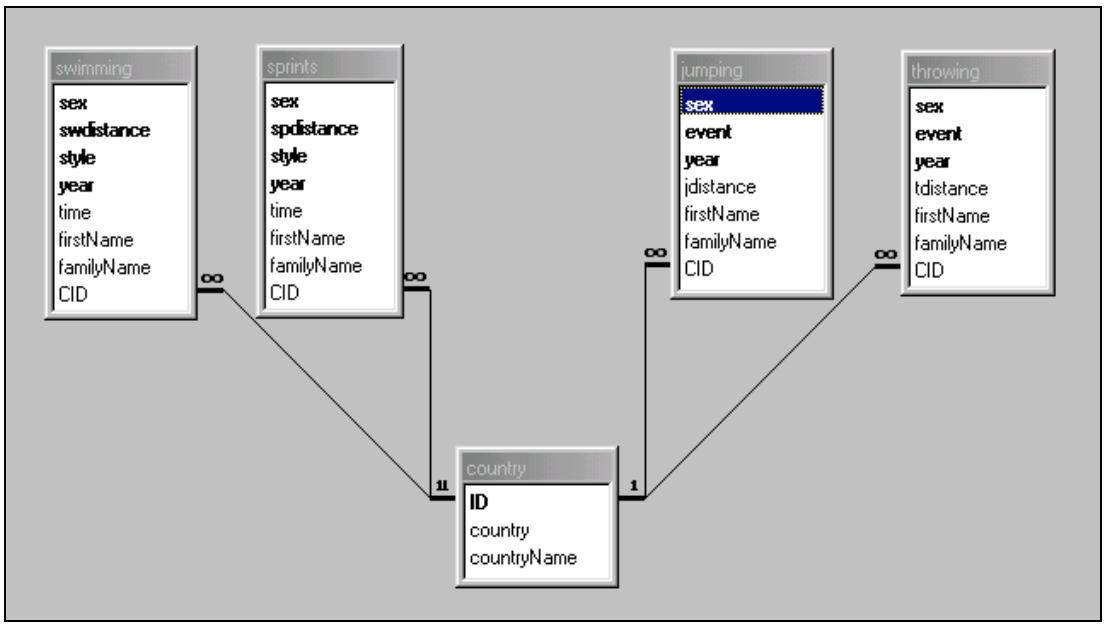


Figure 3.1: Structure of database

There are five tables in the database, with the information about the winners' country in one table (**country**) and the winners' names and numerically coded data containing the results for the various events in four other tables (**swimming**, **sprints**, **jumping** and **throwing**, respectively). These tables have the following fields.

swimming table

sex : 1=male , 2=female
swdisdance : distance in meters
style : 1=freestyle , 2=backstoke , 3=breaststoke , 4=butterfly , 5=medley
year : 1928-2000
time: winning time in seconds
firstname : firstname of the Olympic swimming winner
familyname : familyname of the Olympic swimming winner
CID : ID of the winner's country

sprints table

sex : 1=male , 2=female
spdisdance : distance in meters
style : 1=running , 2=hundles
year : 1928-2000
time: winning time in seconds
firstname : firstname of the Olympic running winner
familyname : familyname of the Olympic running winner
CID : ID of the winner's country

jumping table

sex : 1=male , 2=female
jdisdance : winning distance in meters
event : 1=high jump , 2=long jump , 3=triple jump
year : 1928-2000
firstname : firstname of the Olympic jumping winner
familyname : familyname of the Olympic jumping winner
CID : ID of the winner's country

throwing table

sex : 1=male , 2=female
tdisdance : winning distance in meters
event : 1=shot put, 2=discus , 3=javelin, 4=hammer
year : 1928-2000
firstname : firstname of the Olympic jumping winner
familyname : familyname of the Olympic jumping winner
CID : ID of the winner's country

3.2 Characteristics of Determinants

Table 3.1 shows the type of styles for Swimming, Running, Jumping and Throwing separated by gender. On average the number of athlete between male and female for swimming and jumping is similar. However, there is no record for the jumping female in the triple jump style. This also shows that the number of the running events in both styles for males is almost twice greater than for females. Next we will look at the descriptive statistics in each type.

Type	Style	Sex	
		Men	Women
Swimming	Freestyle	64	56
	Backstroke	26	26
	Breaststroke	25	26
	Butterfly	21	21
	Medley	17	17
Running	Sprint	85	53
	Hurdles	34	16
Jumping	High jump	17	17
	Long jump	17	14
	Triple jump	17	2
Throwing	Shot put	17	14
	Discus	17	17
	Javarin	17	16
	Hammer	17	-

Table 3.1: The styles of four sports for sex

In the next section we begin the preliminary analysis of the data, beginning with the swimming events.

3.3 Description of Swimming Data

To facilitate comparisons between different events, the responses need to be reexpressed as speeds rather than time. This is easily done using a database query involving a calculated field of the form

Speed = [distance]/[time]

The events are coded using the formula

EventID = [swdistance] + 10*[sex] + [style]

This gives a unique identifier for each of the 26 swimming events. The query then produces a table comprising three fields: Speed, Year, and EventID.

Figure 3.2 gives numerical summaries and histograms of the speeds and times of occurrence in all the swimming events recorded in the database.

Variable	Size	Mean	StDev	Skew	Kurt	Min	Med	Max
speed	299	1.584	0.216	0.331	0.548	1.038	1.575	2.282
year	299	1974.930	19.434	-0.792	-0.088	1928.000	1980.000	2000.000
eventID	299	305.903	331.746	2.791	7.334	61.000	214.000	1511.000

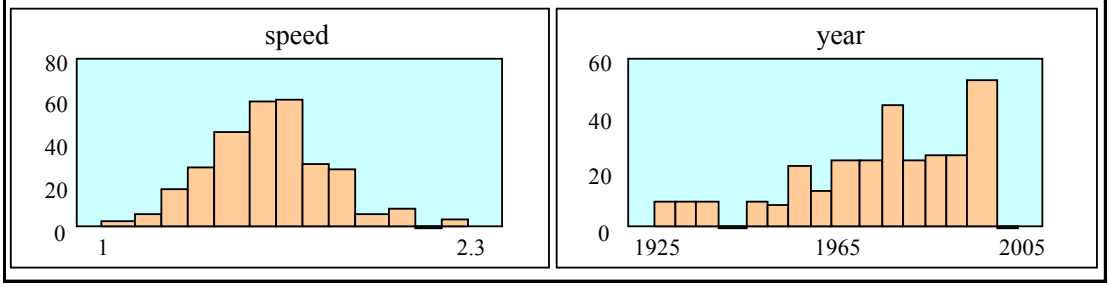


Figure 3.2: Numerical summaries of swimming data with histograms of speeds (left) and years (right)

The speeds range from a minimum of 1.04 to a maximum of 2.28 meters/second, with mean 1.58 and standard deviation 0.22, and the distribution is unimodal and approximately symmetric. The period covers the range 1928 to 2000 and the distribution is skewed to the left, showing that there were more swimming events in the more recent years.

Figure 3.3 shows box plots summarising the distributions of the speeds classified by distance and gender. We see that the 50 meters swimming event has a much higher speed and the 1500 meters event has a lower speed than other distances in men's swimming. The results are similar for women's events, with the 50 meters event faster. However the speeds in the other women's events are similar.

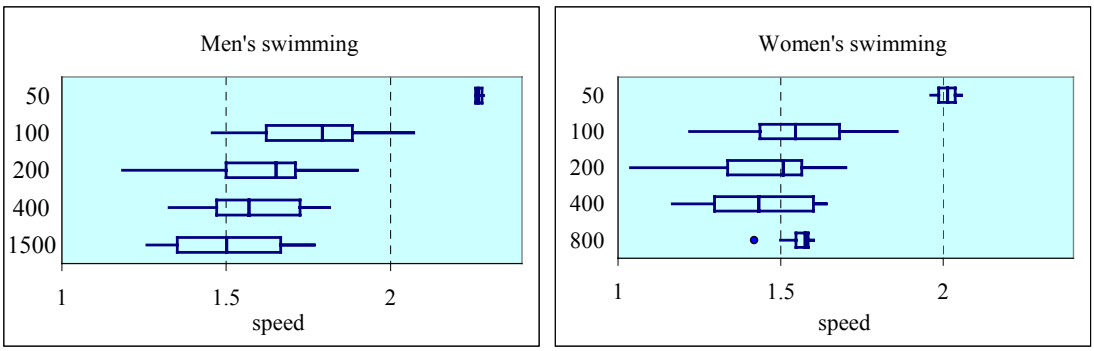


Figure 3.3 Box plots of swimming speeds by distance for men and women

Figure 3.4 shows a comparison of the swimming speeds by year for both men and women. This graph shows how the speeds increased from 1928 to 2000.

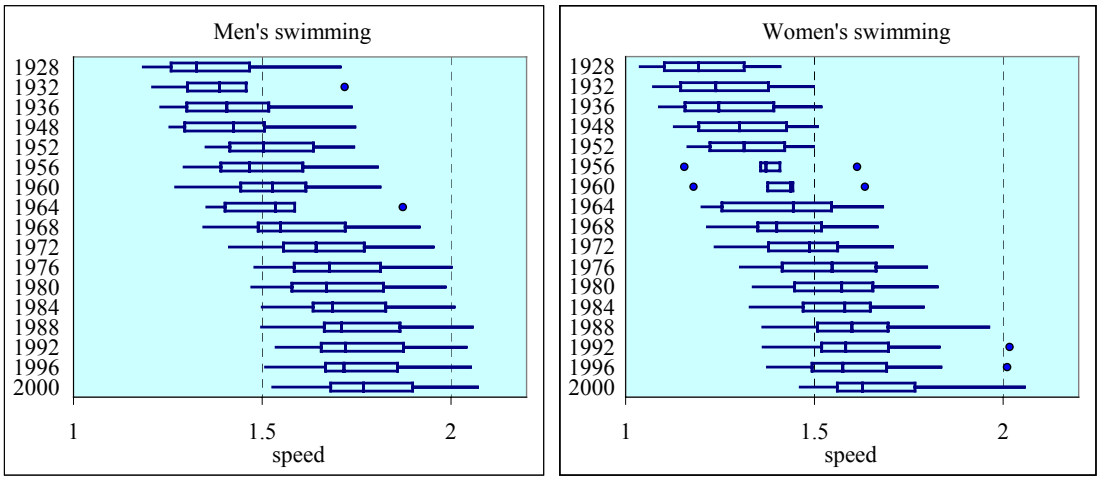


Figure 3.4 Box plots of swimming speed by years for men and women

Figure 3.5 shows a similar comparison of the speeds with respect to both the distance and the type of event.

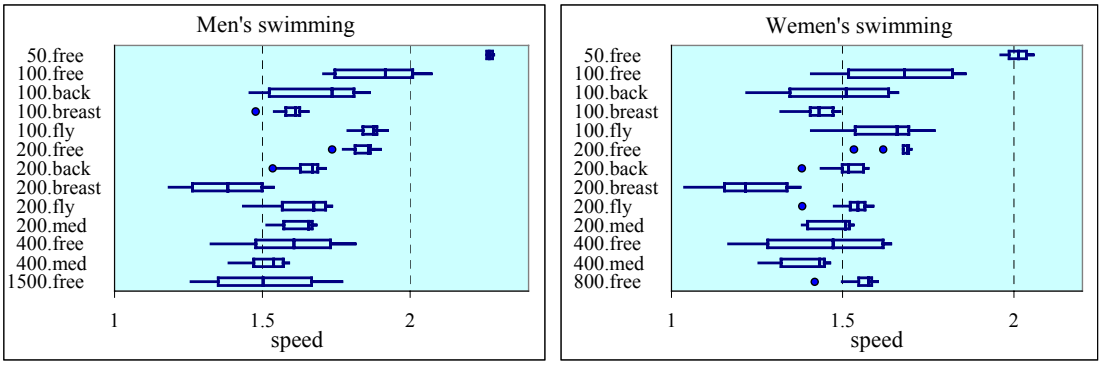


Figure 3.5: Box plots of swimming speed by type of event for men and women

Clearly, the fastest speeds (close to 2.2 meters/second) are achieved in the men's 50 meters freestyle event. The patterns are similar for the men and the women, with the men's speeds ranging from 1.2 to 2.2 meters/second, and the women's speeds about 0.2 meters/second slower, ranging from 1.0 to 2.0 meters/second. In each case the 100 meters freestyle, the 100 meters butterfly and the 200 meters freestyle events all have approximately the same means, although the 100 meters freestyle speeds cover a greater range. In each sex, the 200 meters breaststroke is the slowest event.

Since the comparison is complicated by the fact that different styles have different speeds, Figure 3.6 shows the comparison by year for the men's and women's freestyle events. The result is similar to that shown in Figure 3.4. There is a lot of variation between the different freestyle events corresponding to the different differences.

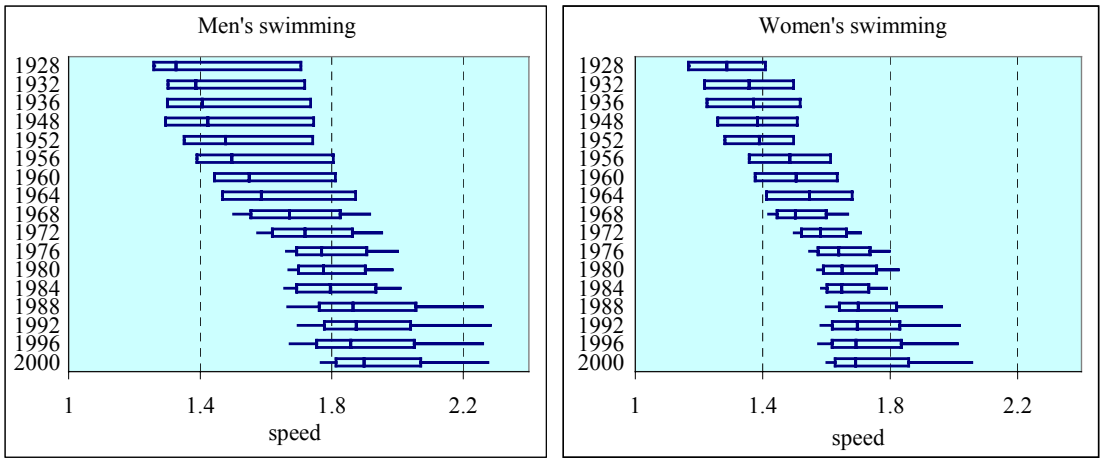


Figure 3.6: Box plots of swimming speed by year for men's and women's freestyle

Figure 3.7 shows the comparison of speed by distance in freestyle events for men and women. As shown in Figure 3.5, the men's 50 meters freestyle has the highest speed and men's 1500 meters freestyle has the lowest speed.

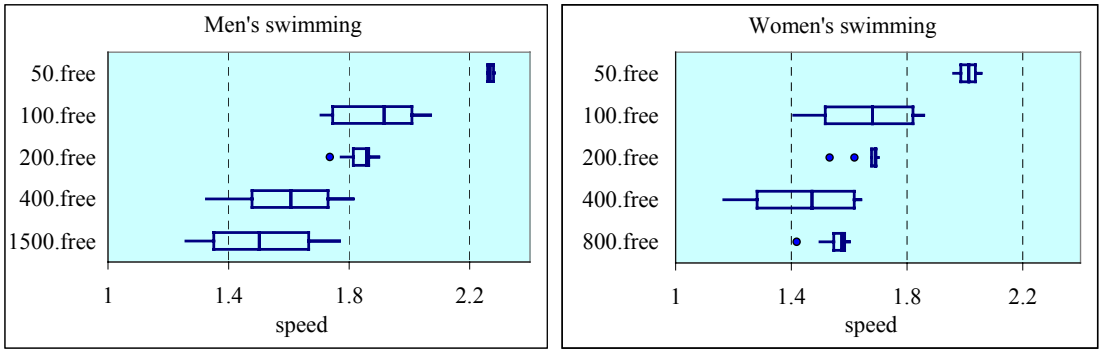


Figure 3.7: Box plots of swimming speed by freestyle events for men and women

Figure 3.8 shows the distribution of the speeds in each year for men's and women's events after adjusting for the 13 different events. As explained in Chapter 2, the adjustment is made by subtracting (or adding) constants corresponding to the different events without changing the sample means, and thus facilitates the comparison of the speed with respect to year. We see the effect of the adjustment in reducing the variation within each year by comparing Figures 3.4 (before adjustment) and 3.8 (after adjustment).

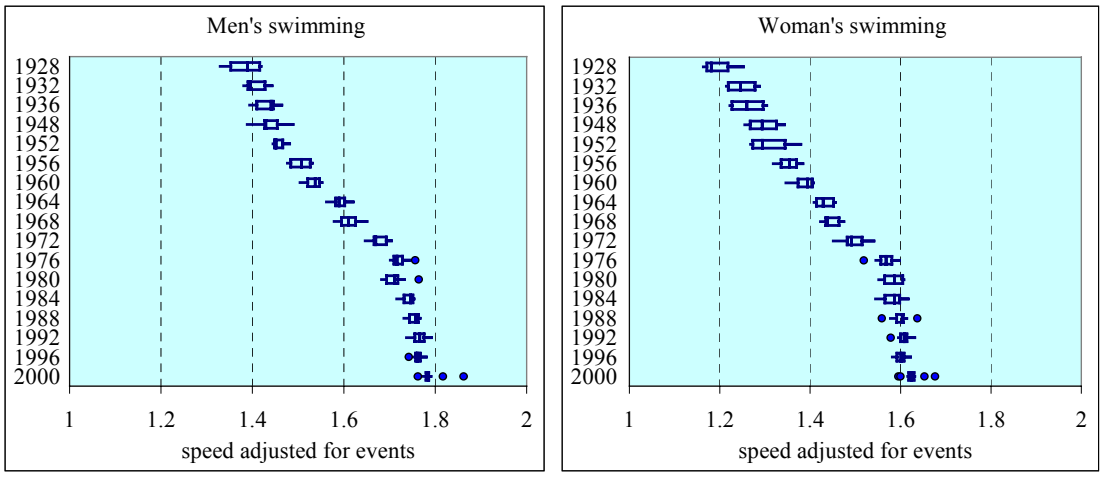


Figure 3.8: Box plots of event-adjusted swimming speeds by year for men and women

While Figure 3.8 shows very clearly how the speeds have increased over the 62-year period, the trend is even clearer when looking at confidence intervals for the means, as shown in Figure 3.9. This graphical method for comparing means is described in Chapter 2.

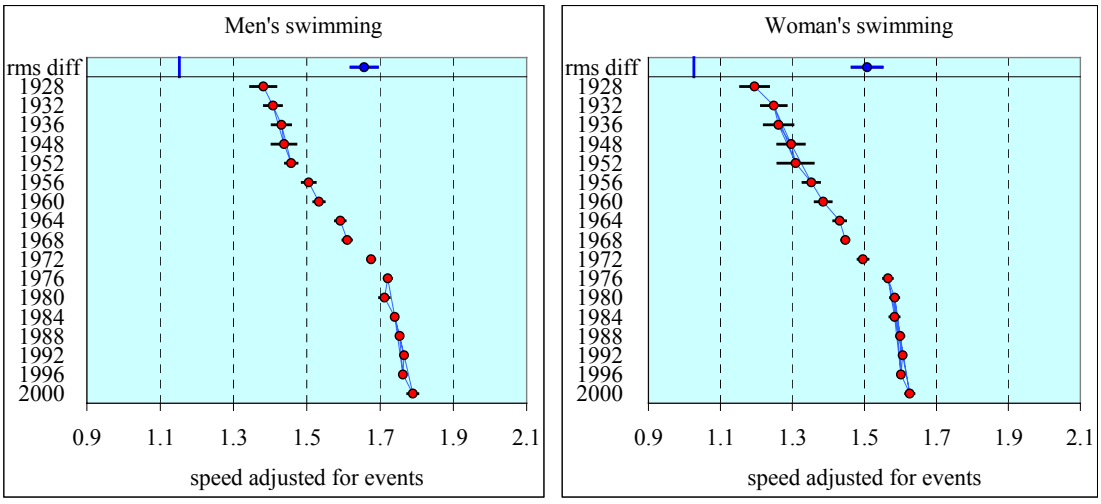


Figure 3.9: 95% confidence intervals of event-adjusted speeds by year for men and women

Figure 3.9 shows that, for each sex, the swimming speeds increased consistently from 1928 to 1968, with even greater increases in 1972 and 1976, but then increased more slowly from 1980 to 2000. The patterns for men and women are very similar, with a sconstant gap of close to 0.2 meters/second separating the men from the women.

The largest residual for the men's events is 0.057 meters/second, arising from the 1500 meters freestyle event at the Moscow 1980 Olympics. Vladimir Salnikov from the Soviet Union won this event in the time of 14 minutes 58.27 seconds. The largest negative residual for the men's events is –0.052 meters/second, corresponding to the time of 5 minutes 1.6 seconds recorded by Victoriano Zurillo from Argentina at the Amsterdam 1928 Olympics. The largest residual for the women's events is 0.070 meters/second, arising from the 200 meters breaststroke event at the Helsinki 1952 Olympics. Eva Szekely from Hungary won this event in the time of 2 minutes 51.77 seconds. The largest negative residual for the women's events is –0.047 meters/second, corresponding to the time of 2 minutes 33,35 seconds recorded by Kosheveya from the Soviet Union at the Montreal 1976 Olympics.
Figure 3.10 shows box plots comparing the speeds in the different events for men and women after adjusting for years.

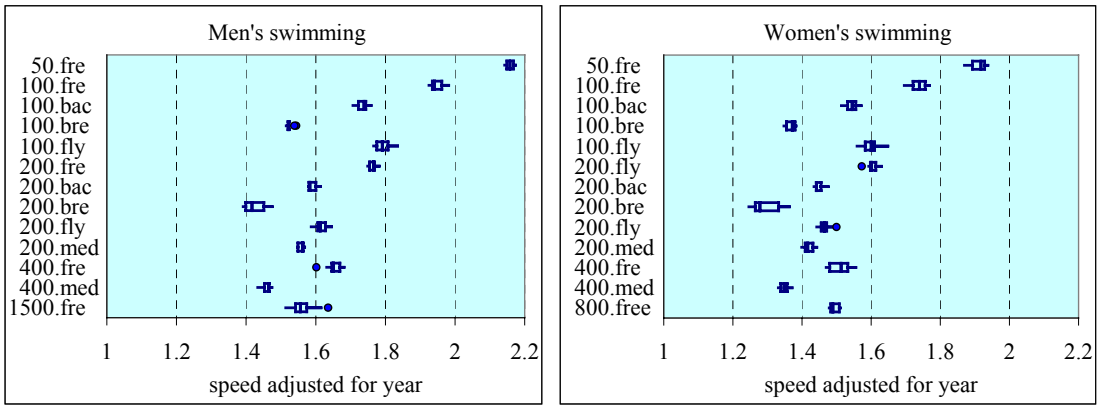


Figure 3.10: Box plots of swimming speed by events men and women adjusted for year

We see the effect of the adjustment in reducing the variation within each event by comparing Figures 3.5 (before adjustment) and 3.10 (after adjustment).

While Figure 3.10 shows very clearly how the speeds depend on the event, this dependence is seen even more clearly in the confidence intervals for the means, as shown in Figure 3.11.

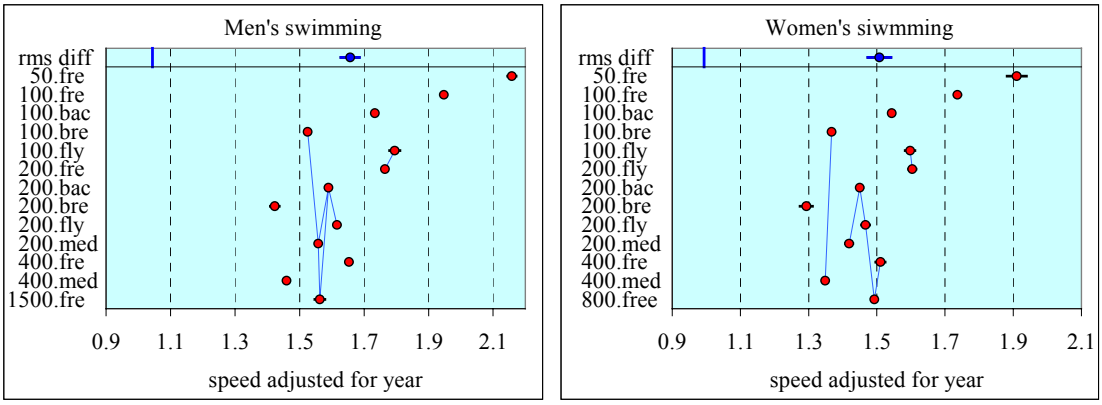


Figure 3.11 Means & 95% confidence intervals of event speeds adjusted for year

The patterns are similar for each sex. In each case the 50 meters freestyle is fastest, followed by the 100 meters freestyle event, which is approximately 0.2 meters/second slower. Next come the 100 meters butterfly and the 200 meters freestyle events, closely followed by the 100 meters backstroke, which are another 0.2 meters/second slower. After that, the patterns differ slightly between the men and the women, but this difference is largely accounted for by the fact that the men swim a 1500 meters freestyle event and the women swim an 800 meter freestyle event instead. In each case the slowest event is the 200 meters breaststroke. There is also some evidence that the gap between the men and the women is larger in the faster events. For the fastest event (the 50 meters freestyle) this gap is more than 0.2 meters/second, while for the slowest event (the 200 meters breaststroke) the gap is only slightly more than 0.1 meters/second.

In the next section we report the results from a similar analysis for the Olympic running speeds.

3.4 Description of Running Data

As in the case of the Olympic swimming events, to facilitate comparisons between different events, the winning Olympic running times need to be reexpressed as speeds. This is done using a database query involving a calculated field of the form

Speed = [spdistance]/[time]

The events are coded using the formula

EventID = [spdistance] + 2*[sex] + [style]

This gives a unique identifier for each of the 13 running events. However, the women's 80 meters hurdles event changed to 100 meter hurdles in 1928-1968. For simplicity, we combine these two events, so there are 12 running events. The query then produces a table comprising three fields: Speed, Year, and EventID.

Figure 3.12 shows numerical summaries and histograms for the winning speeds and the years of occurrence.

Variable	Size	Mean	StDev	Skew	Kurt	Min	Med	Max
speed	188	8.352	1.032	-0.083	-0.863	5.848	8.390	10.352
year	188	1968.745	21.098	-0.360	-0.857	1928	1972	2000
EventID	188	426.181	415.141	1.531	1.454	103.000	205.000	1503.000

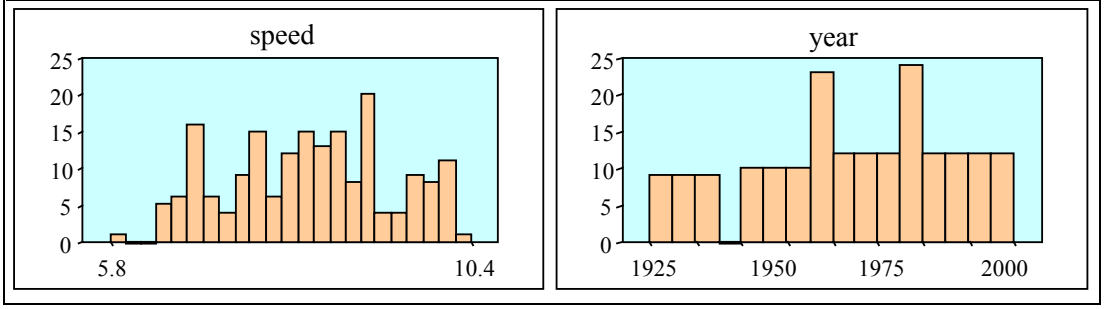


Figure 3.12: Numerical summaries of running data with histograms of speeds (left) and years (right)

The speeds range from 5.85 to 10.35 meters/second, and their distribution is irregular. The years of occurrence range from 1928 to 1996, with an approximately uniform distribution, indicating that the number of running events has not varied greatly over the years.

Figure 3.13 shows box plots summarising the distributions of the speeds by type of event – sprints and hurdles. It shows that the median speeds are higher in the sprints than in the hurdles for both sexes, but there is quite a lot of variation.

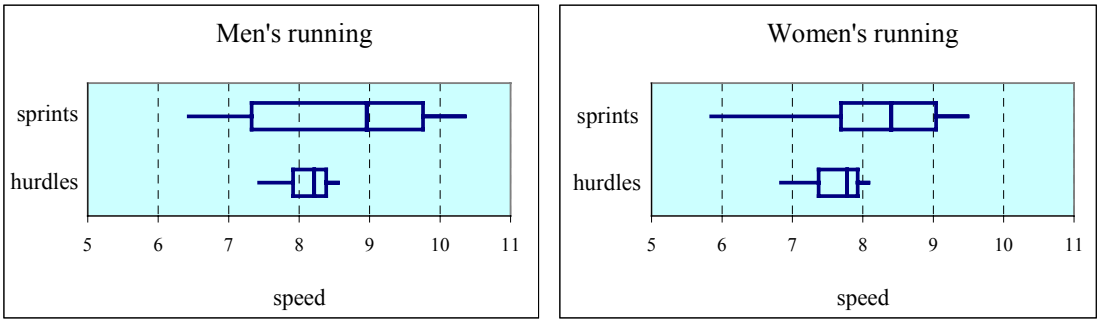


Figure 3.13: Box plots of running speeds by type of event for men and women

The variation in running speed by year is shown in Figure 3.14. There is so much variation within each year that it is difficult to see any changes from year to year.

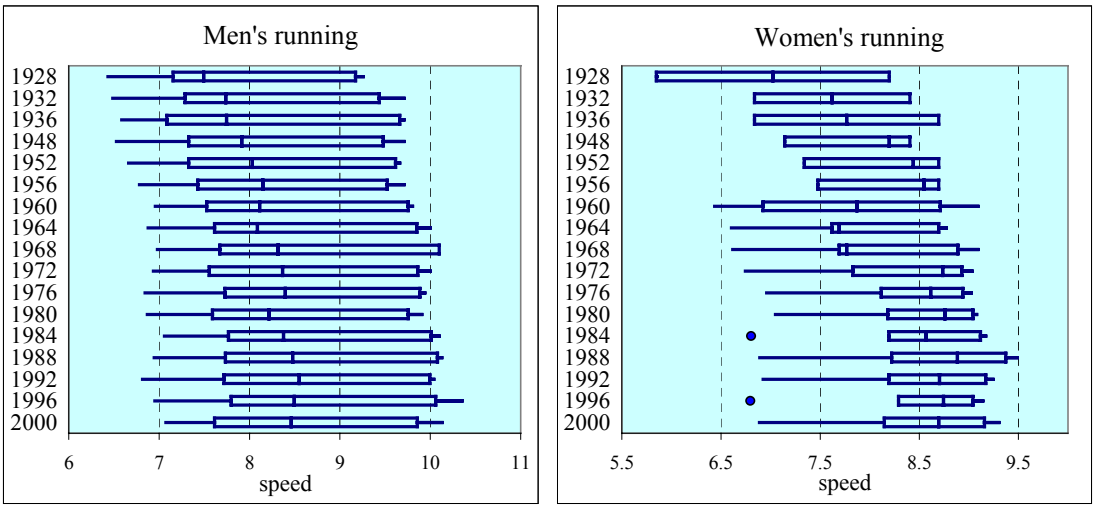


Figure 3.14: Box plots of running speed by year for men and women

Figure 3.15 compares the speeds for the various distances, and this shows that the speeds in the shorter distances are much greater than those in the longer distances. The graph also shows how much slower the hurdles events are than the corresponding sprint events.

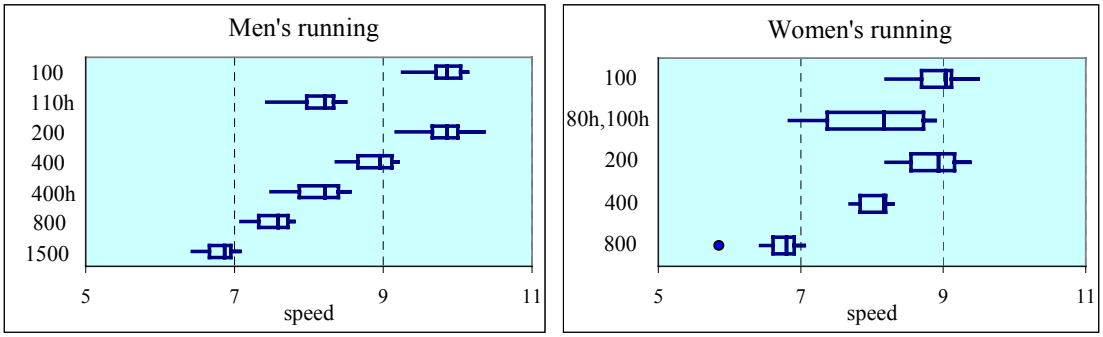


Figure 3.15: Box plots of running speeds by event for men and women

Although men have competed in two more events than women (the 400 meters hurdles and the 1500 meters sprint), the patterns are similar for the two sexes, with the men approximately 1 meter/second faster than the women in each event. In each case the 100 and 200 meters sprint has the highest speed and the speeds decrease with distance, by approximately 1 meter/second at each doubling of the distance, in the sprinting events. And in each case the shorter hurdles event has a speed that is almost 2 meters/second slower than the sprint event. However, the difference in speed

between the sprint and the hurdles for the men's 400 meters is only about 1 meter/second. The slowest speed occurred in the women's 800 meters sprint. This corresponds to a winning time of 2 minutes 16.8 seconds, by Lina Radke from Germany in the 1928 Amsterdam Olympics. This event from discontinued until the 1960 Rome Olympics, when Ludmilla Shevtsova from the Soviet Union won the event in 2 minutes 4.3 seconds.

Figure 3.16 shows the distributions of the speeds for each sex after adjusting for the type of event. This graph should be compared with Figure 3.14, which shows the same data without the adjustment. The trends can now be seen quite clearly.

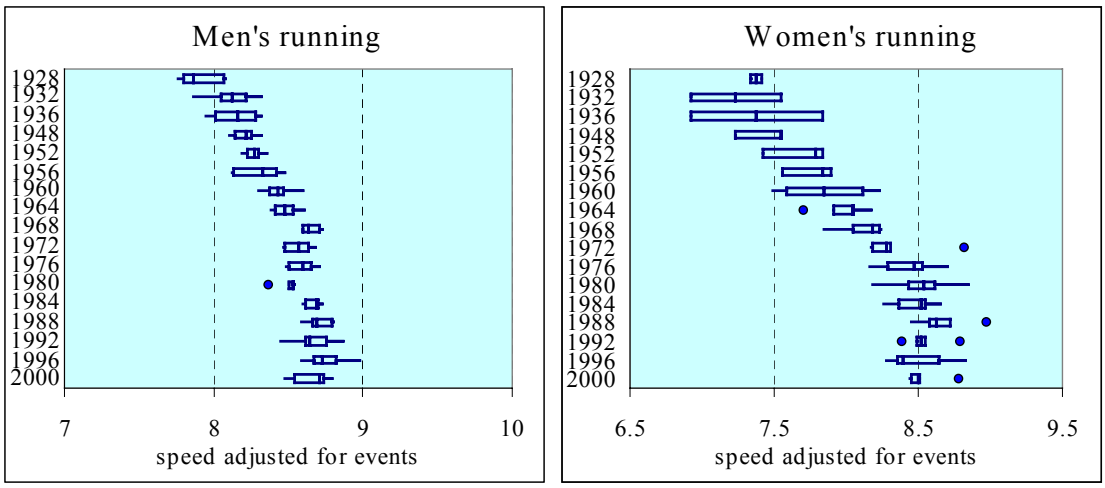


Figure 3.16: Box plots of running speed by year for men and women adjusted for events

These trends can be seen even more clearly in Figure 3.17, which shows 95% confidence intervals of the mean speeds after adjusting for the differences between the events. While the speeds increased with time for both men and women, the patterns are different. For the men, there was a substantial rate of increase from the Amsterdam 1928 Olympics to the Mexico City 1968 Olympics, but not so much change between 1968 and 1996. The rate of increase in the women's speeds during the period from 1928 to 1968 was even greater than that for mean, and continued until 1988, after which time it decreased. On average, the men ran approximately 0.5 meters/second faster than the women.

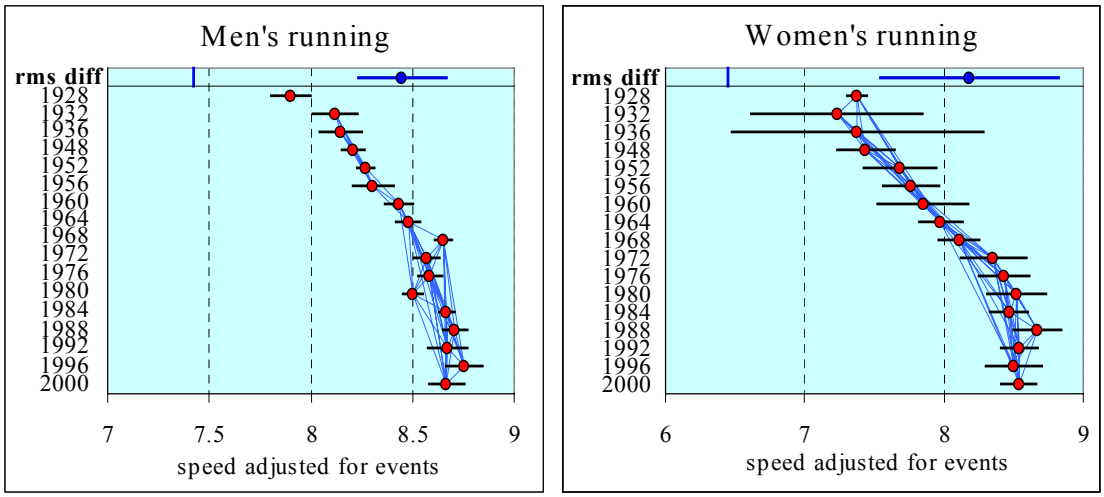


Figure 3.17: Means & 95% confidence intervals adjusted for events

Figure 3.18 shows the speeds in the various events for men and for women after adjusting for the effect of year. The highest speed is in the 100 meters and the lowest speed in the 1500 meters for men's running. In this figure the speed of running by events gives the highest speed in the 100 meters and the lowest speed in the 800 meters for women's running.

Figure 3.19 shows the 95% confidence intervals for these data.

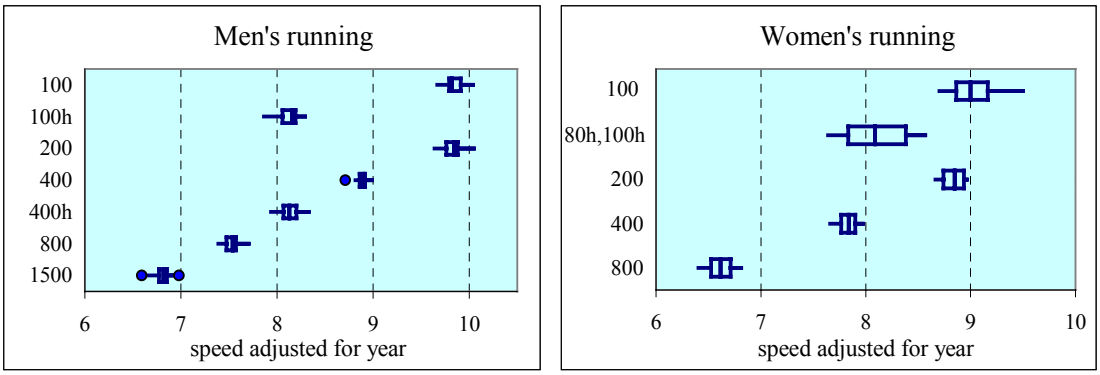


Figure 3.18: Box plots of running speed by events for men and women adjusted for year

In Figure 3.19 the means of speed give the lowest of speed in the women's 800 meters and the men's 1500 meters and the highest in the men's 100 and 200 meters.

It is interesting to note that the speeds in the men's 110 meters and 400 meters hurdles events are the same.

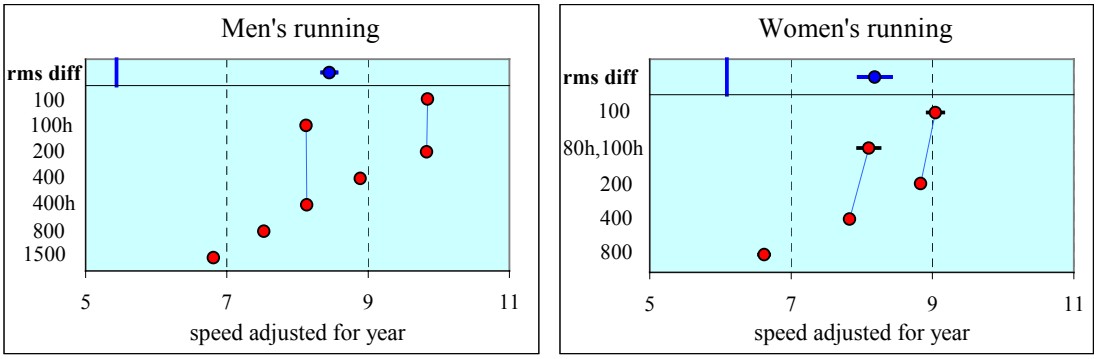


Figure 3.19: Means & 95% confidence intervals adjusted for year

3.5 Description of Jumping Data

In the jumping and throwing events the response is distance, not time as in the swimming and running events. Since the distances vary greatly in the three events, from less than 2 meters in the high jump to almost 20 meters in the triple jump, and the variation increases with distance, some transformation of the data is desirable. Figure 3.20 shows numerical summaries and histograms of the distances.

Numerical Summaries: jumping

Variable	Size	Mean	StDev	Skew	Kurt	Min	Med	Max
event	84	3	1.456	0.120	-1.118	1	3	6
year	84	1968.57	21.448	-0.302	-0.922	1928	1970	2000
distance	84	7.385	5.671	0.717	-0.849	1.590	6.890	18.170
log distance	84	0.722	0.374	-0.018	-1.543	0.201	0.838	1.259

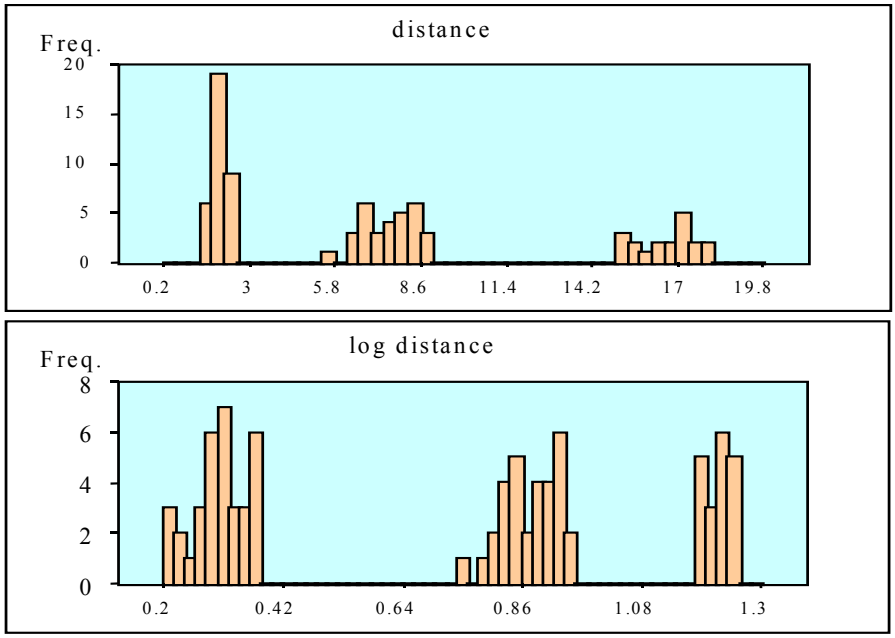


Figure 3.20: Numerical summaries of jumping data with histograms of distance (+ log-transformed)

Since the distances vary greatly in the three events, from less than 2 meters in the high jump to almost 20 meters in the triple jump, and the variation increases with distance, some transformation of the data is desirable. The histograms in Figure 3.20 show that the variation increases substantially as the distance increases, but the variation is more stable after taking logarithms.

Figure 3.21 compares the box plots of the winning distances, and shows that the standard deviation of distance is highest for the men's triple jump (0.91) and lowest for the women's high jump (0.16). The women's triple jump, with only two results (1996 and 2000) is omitted.

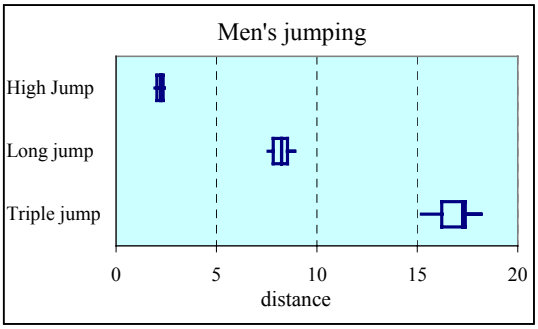
Figure 3.22 shows the same graphs after the log-transformation.

Comparison: distance

event	Size	Mean	SE	StDev
1	17	2.195	0.038	0.157
2	17	8.226	0.099	0.410
3	17	16.869	0.220	0.908

Comparison: distance

event	Size	Mean	SE	StDev
1	17	1.851	0.039	0.163
2	14	6.743	0.120	0.448



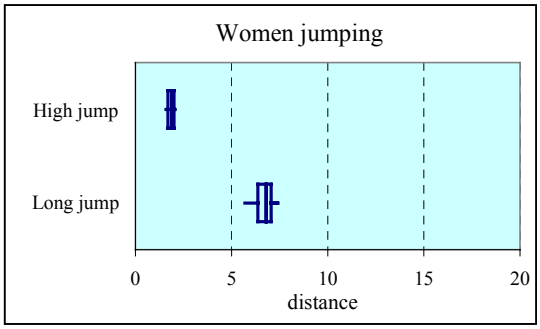


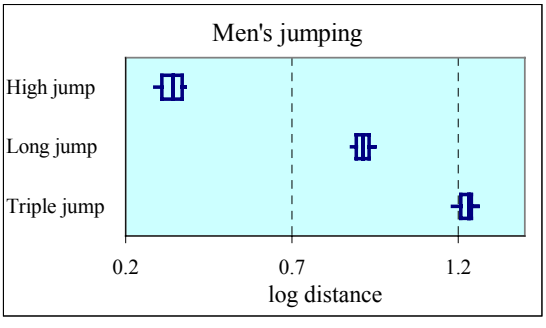
Figure 3.21: Box plots of jumping distance by events for men and women

Comparison: log distance

event	Size	Mean	SE	StDev
1	17	0.340	0.008	0.031
2	17	0.915	0.005	0.022
3	17	1.226	0.006	0.024

Comparison: log distance

event	Size	Mean	SE	StDev
1	17	0.266	0.009	0.039
2	14	0.828	0.008	0.030



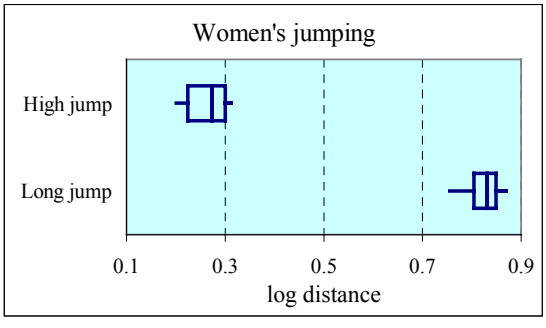
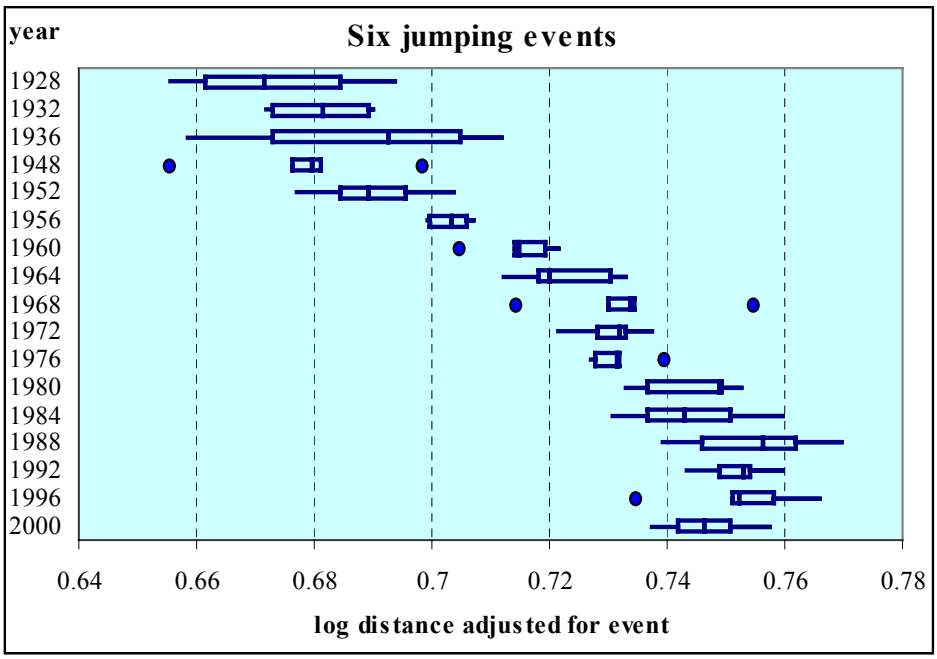


Figure 3.22: Box plots of jumping log distance by events for men and women

Now the standard deviations are similar, with the highest for the women's high jump (0.039) and the lowest for the men's long jump (0.022). This indicates that taking logarithms makes the statistical assumption of variance homogeneity reasonable.

Figure 3.23 shows how the log-transformed winning distances increased with year. In this plot, the data for the women's triple jump are included. The distributions are adjusted for differences between the six events using two-way anova.



3.23: Box plots of jumping log distance by year adjusted for six events

Figure 3.24 shows the corresponding 95% confidence intervals. The trend shown in Figure 3.24 is for the performances to improve substantially with time from 1928 to 1968, with a slowing down of this trend in the more recent years.

Figure 3.25 shows box plots of the data for the six events after making an adjustment for the years, and Figure 3.26 shows the corresponding 95% confidence intervals. These intervals are so short relative to the differences between the events that they cannot be seen on the plot.

The value of logarithmic distance in woman high jump is lowest and the value of log distance in men's triple jump is highest. The mean of log distance adjusted for year in men's triple jump is highest and in women's high jump is lowest. The pattern shows that for both the high jump and the long jump, the mean for the men is slightly greater

(by approximately 0.08) than the mean for the women. This is on the base 10 log scale, and translates to a proportionate change of 10^0.08 = 1.202, that is, a 20% change.

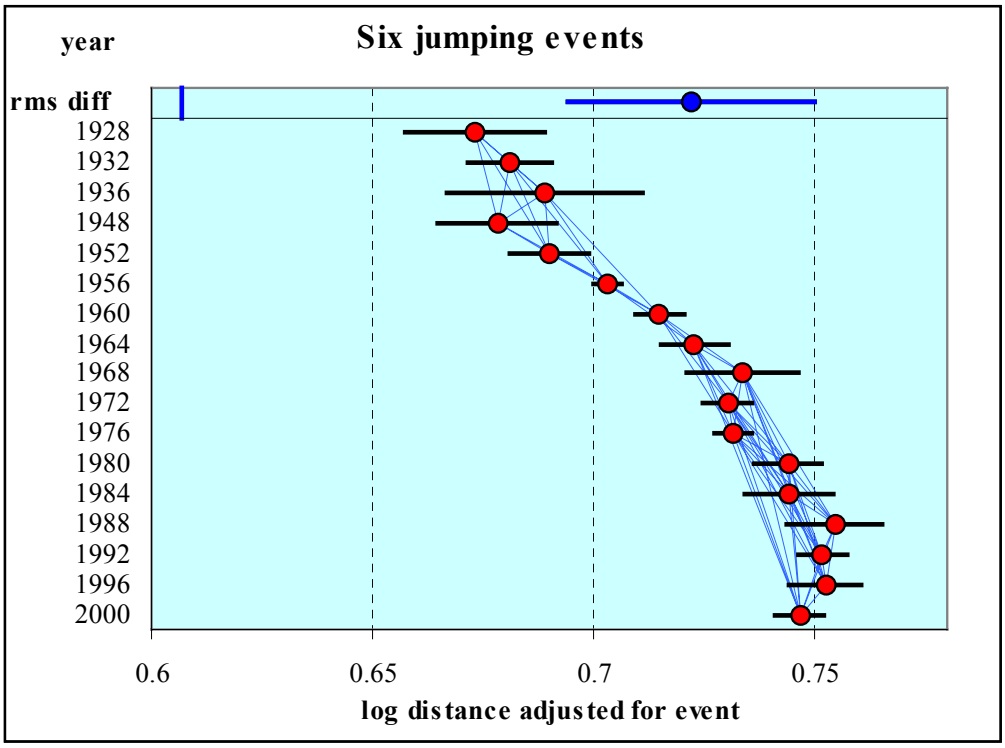


Figure 3.24: Means & 95% confidence intervals of log-distance adjusted for year

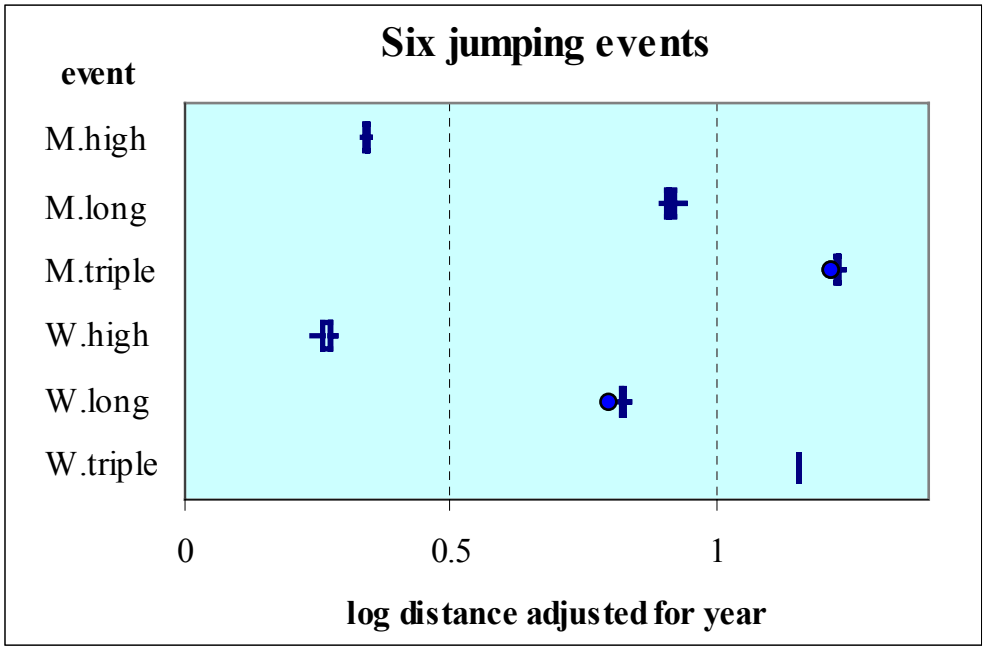


Figure 3.25: Box plots of jumping log-distance by event adjusted for year

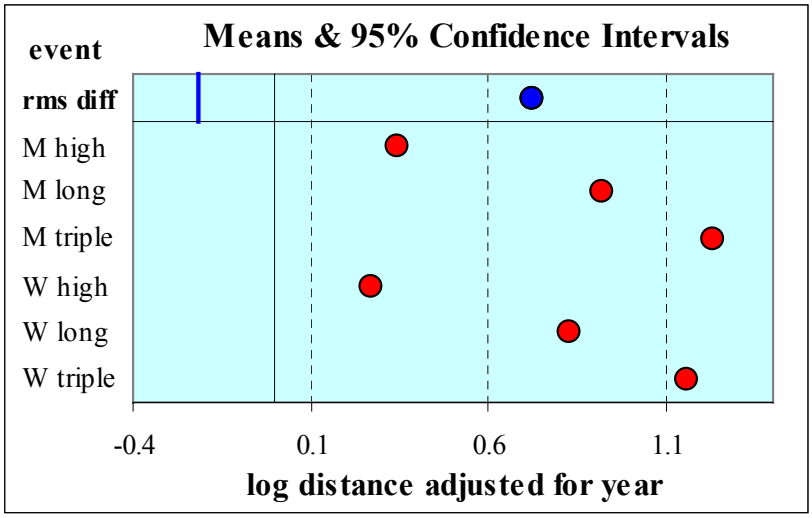


Figure 3.26: Means & 95% confidence intervals

3.6 Description of Throwing Data

In the jumping and throwing events the response is distance, not time as in the swimming and running events. Since the mean distances vary greatly in the four events, from just over 19 meters in the shot put to over 70 meters in the javelin and hammer throws, and the variation increases with distance, some transformation of the data is desirable. Figure 3.27 shows numerical summaries of the distances before and after taking base 10 logarithms. The winning distances range from 13.75 meters to 94.58 meters (a factor of seven from lowest to highest), whereas the range of the log-transformed distances is 1.14 to 1.98 (a factor of less than 2).

Numerical Summaries: Throwing

Variable	Size	Mean	StDev	Skew	Kurt	Min	Med	Max
year	115	1967.83	21.135	-0.274	-0.929	1928	1968	2000
distance	115	53.955	23.967	-0.347	-1.158	13.750	60.340	94.580
log distance	115	1.672	0.250	-0.801	-0.927	1.138	1.781	1.976

Figure 3.27:Numerical summaries of throwing data, with log-transformed distance

Figure 3.28 shows numerical summaries and box plots of the winning distances for each event for the two sexes, and shows that the mean winning distance is highest for the men's javelin (72.0 meters) and lowest for the women's shot put (19.30 meters).

Comparison: distance for men's throwing

event	Size	Mean	SE	StDev
1=Shot put	31	19.475	0.433	2.411
2=Discus	34	59.737	1.637	9.544
3=Javelin	33	72.015	2.578	14.812
4=Hammer	17	70.206	2.737	11.284

Comparison: distance for women's throwing

event	Size	Mean	SE	StDev
1=Shot put	14	19.299	0.711	2.662
2=Discus	17	58.639	2.751	11.341
3=Javelin	16	60.212	2.482	9.928

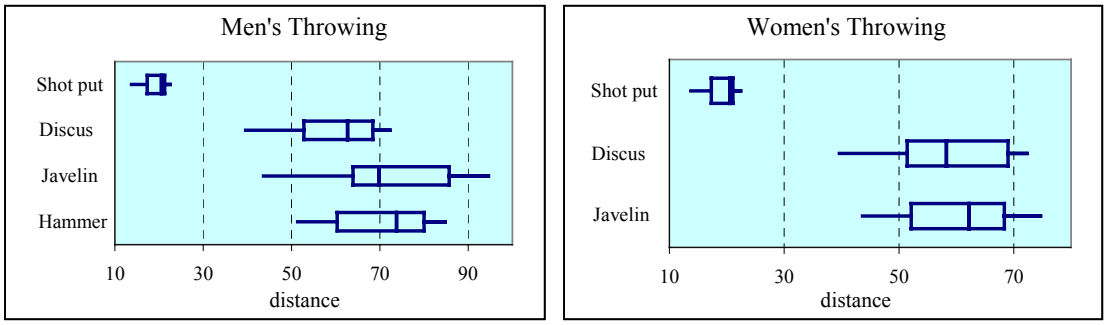


Figure 3.28: Box plots of throwing distances by event for men and women

Note that the standard deviation is also highest for the men's javelin throw (14.81 meters) and lowest for the men's shot put (2.41). The events with higher mean distances also have higher standard deviations, and the highest standard deviation is six times the lowest standard deviation, which means that the statistical assumption that the standard deviations be the same is invalid.

Figure 3.29 shows the same comparison after log-transforming the distances. Again, the mean of log distance is highest for the men's javelin and lowest for the women's shot put. But now the standard deviations are more evenly distributed. The highest standard deviation is 0.095 (again for the men's javelin throw) but the lowest is 0.057 (again for the men's shot put), and this ratio is only 1.6. However, while the transformation has made the spreads more uniform, some negative skewness is now evident in the box plots.

Since the statistical assumption of equal variances is now more nearly satisfied, further analysis is based on the log-transformed distances.

Figure 3.30 shows how the log-transformed winning distances increased with year. In this plot, these data are adjusted for the seven events.

Comparison: log distance for men's throwing

event	Size	Mean	SE	StDev
1=Shotput	31	1.286	0.010	0.057
2=discus	34	1.770	0.013	0.074
3=javelin	33	1.848	0.017	0.095
4=hammer	17	1.841	0.018	0.073

Comparison: log distance women's throwing

event	Size	Mean	SE	StDev
1=Shotput	14	1.281	0.017	0.064
2=discus	17	1.760	0.022	0.090
3=javelin	16	1.774	0.019	0.075

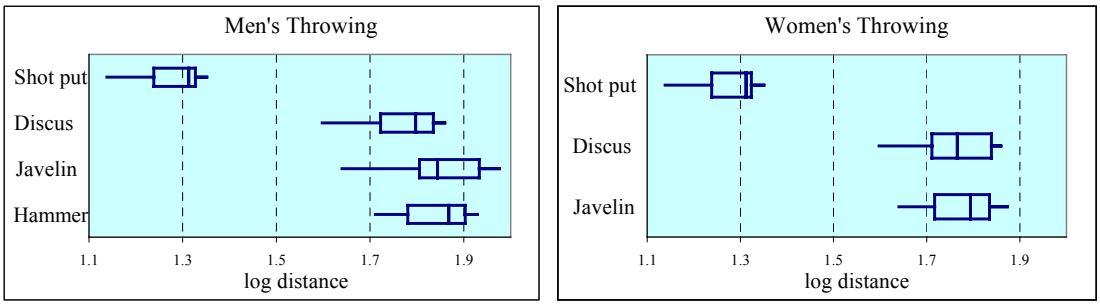


Figure 3.29: Box plots of throwing log distance by events for men and women

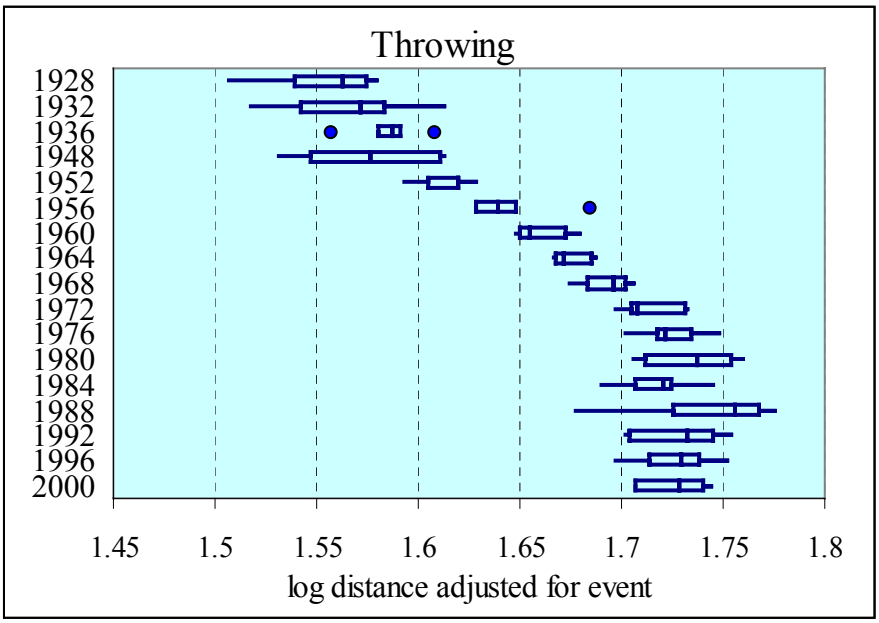


Figure 3.30: Box plots of throwing log-distance by year adjusted for events

Figure 3.31 shows the corresponding 95% confidence intervals. The trend shown in figure 3.31 is for the performances to improve substantially with time from 1928 to 1968, with leveling off of this trend in the more recent years.

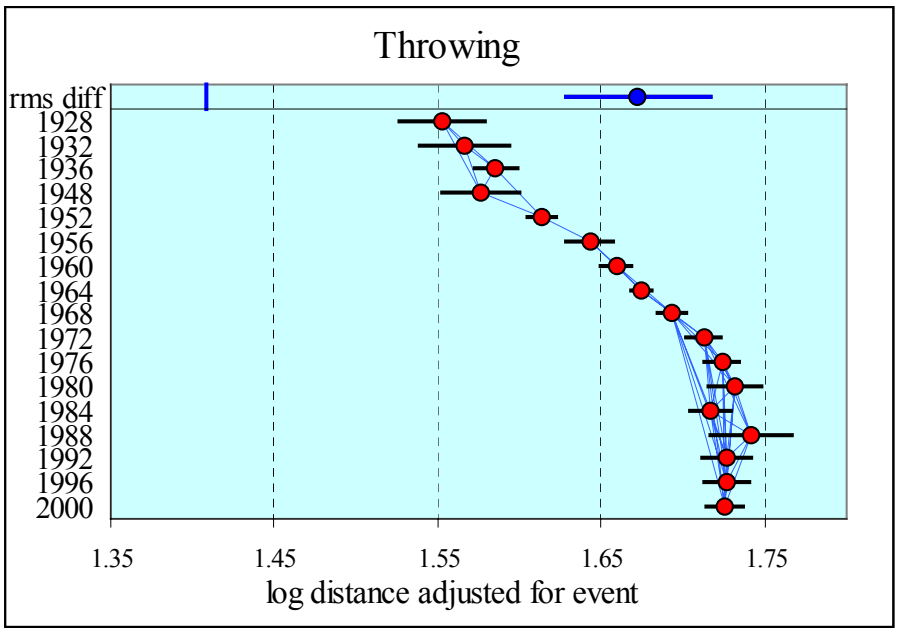


Figure 3.31: Means & 95% confidence intervals of log-distance adjusted for events

Figure 3.32 shows box plots of the data for the seven events after making an adjustment for the years, and Figure 3.33 shows the corresponding 95% confidence intervals. These intervals are so short relative to the differences between the events that that cannot be seen on the plot. The value of log distance in woman shot put is lowest and the value of log distance in men's javelin is highest. The mean of log distance adjusted for year in men's javelin is highest and in women's shot put is lowest.

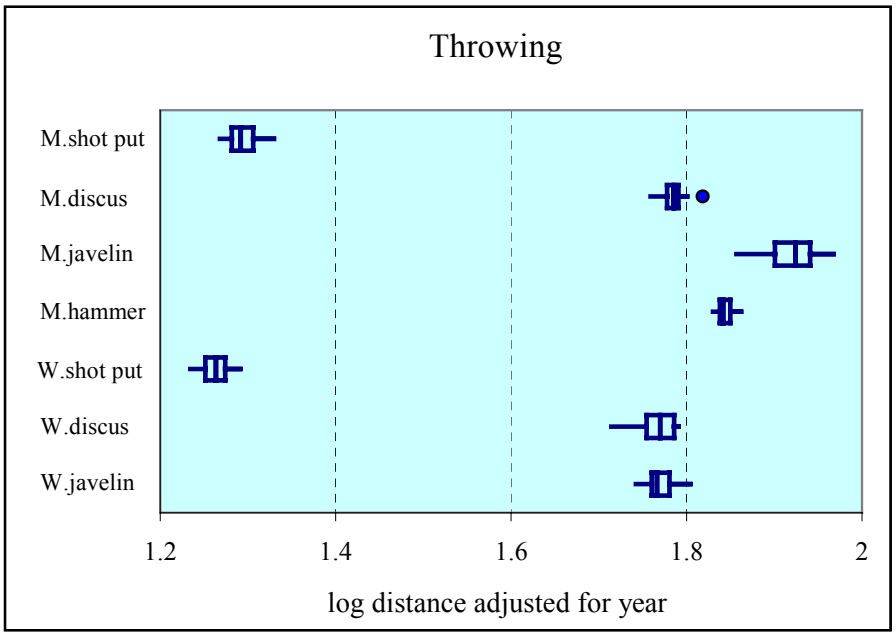


Figure 3.32: Box plots of throwing log-distance by event adjusted for year

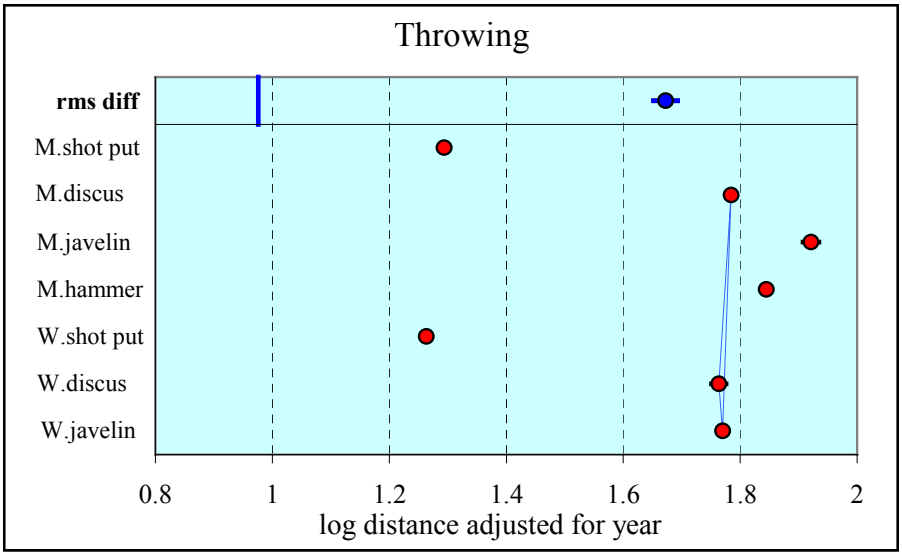


Figure 3.33: Means & 95% confidence intervals

3.7 Summary

The preliminary results may be summarised as follows:

For the swimming and sprints events, the results are appropriately expressed as speeds. After adjusting for differences between events, the winning speeds show an increasing trend for both men and women. In each case this trend slows down towards the end of the 20th century, and the trend may not be the same for men and women.

For the jumping and throwing events, the results are appropriately expressed as logarithms of the winning distances. After adjusting for differences between events, and combining the results for the men's and women's events, the log-transformed winning distances show an increasing trend. For the jumping events, this trend slows down towards the end of the 20th century. For the throwing events, the trend increases slowly at first, then more rapidly, and finally levels off towards the end of the 20th century.

In the next chapter we develop a predictive model for these performances.