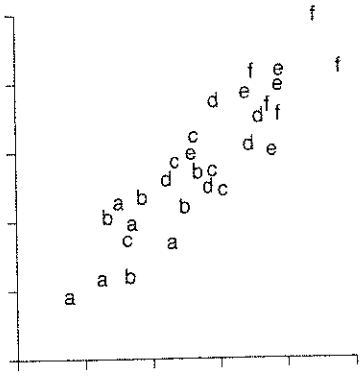
Exercise Set B

1. In the figure below, 6 scatter diagrams are plotted on the same pair of axes; in the first, the points are marked "a"; in the second, "b"; and so forth. For each of the 6 diagrams taken on its own, the correlation is around 0.6. Now take all the points together. For the combined diagram, is the correlation around 0.0, 0.6, or 0.9?



4. ECOLOGICAL CORRELATIONS

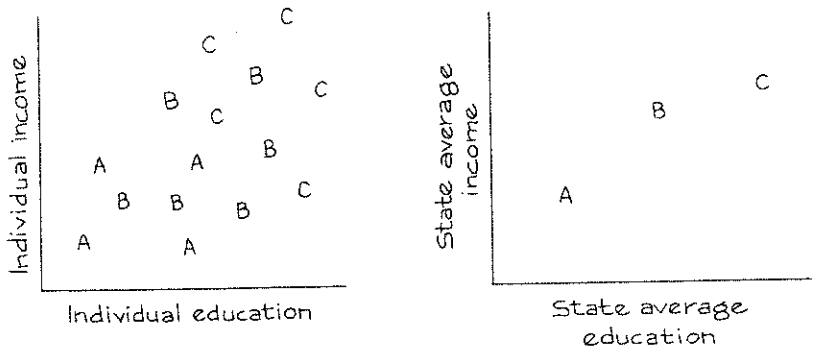
In 1955, Sir Richard Doll published a landmark article on the relationship between cigarette smoking and lung cancer.[2] One piece of evidence was a scatter diagram showing the relationship between the rate of cigarette smoking (per capita) and the rate of deaths from lung cancer in eleven countries. The correla-

tion between these eleven pairs of rates was 0.7, and this was taken as showing the strength of the relationship between smoking and cancer. However, it is not countries which smoke and get cancer, but people. To measure the strength of the relationship for people, it is necessary to have data relating smoking and cancer for individuals rather than countries. Such studies are available, and show that smoking does indeed cause cancer.

The statistical point: correlations based on rates or averages can be misleading. Here is another example. From Current Population Survey data for 2005, you can compute the correlation between income and education for men age 25–64 in the United States: $r \approx 0.42$. For each state (and D.C.), you can compute average educational level and average income. Finally, you can compute the correlation between the 51 pairs of averages: $r \approx 0.70$. If you used the correlation for the states to estimate the correlation for the individuals, you would be way off. The reason is that within each state, there is a lot of spread around the averages. Replacing the states by their averages eliminates the spread, and gives a misleading impression of tight clustering. Figure 6 shows the effect for three states.[3]

> *Ecological* correlations are based on rates or averages. They are often used in political science and sociology. And they tend to overstate the strength of an association. So watch out.

Figure 6. Ecological correlations (based on rates or averages) are usually too big. The panel on the left represents income and education for individuals in three states, labeled A, B, C. Each individual is marked by a letter showing state of residence. The correlation is moderate. The panel on the right shows the averages for each state. The correlation between the averages is almost 1.



Exercise Set D

1. The table at the top of the next page is adapted from Doll and shows per capita consumption of cigarettes in various countries in 1930, and the death rates from lung cancer for men in 1950. (In 1930, hardly any women smoked; and a long period of time is needed for the effects of smoking to show up.)