

Berkson's fallacy

The Canadian immigration system uses a points system to admit new immigrants. Let's work with an inspired by real life but greatly simplified version of it that gives a potential immigrant 0–100 points for their employment prospects (education, job skills, language skills, whether their field is in demand or not) and 0–100 points for their family connections already in Canada (more for a spouse, children; fewer for grandparents or grandchildren, etc.).

The following [list contains 100 potential immigrants](#).

1. The job points and family points were completely randomly generated and are totally independent. Explain what that means.
2. Highlight the columns for job skills and family size, then go to Insert, then Chart. Under chart type, select Scatter chart. Copy your Scatter chart below and explain what it means.
3. Fake Canada only admits half of the immigrants. Select all three columns. Go to Data, then Sort range, then Advanced range sorting options. You'll want to select Data has a header row, then sort by Points Z->A.
 - A. Delete the lowest half of the list (everyone below 100 points).
 - B. Repeat step 2. Copy your Scatter chart below and explain how it differs from step 2. What does it mean?
4. Sort your list of remaining immigrants by job skill Z->A. For the top ten immigrants by job skills, find the average...
 - A. Job skill score (this should be close to 100)
 - B. Family size score

5. Sort your list of remaining immigrants by family size Z->A. For the top ten immigrants by job skills, find the average...

A. Job skill score

B. Family size score (this should be close to 100)

6. Is it the case, for newly admitted immigrants, that a higher job skill score allows you to predict a lower family size score? And vice versa? Explain.

7. For the pool of all potential immigrants, were the two scores correlated? Explain.

8. Explain this spurious correlation in a way a non-statistician would understand.