Ch. 8, 9: Correlation

- Correlation coefficient r, where:

$$r = \frac{\sum_{i=1}^{n} z_{x_i} z_{y_i}}{n}$$

- Another formula:

$$\frac{(x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 (y - \bar{y})^2)}}$$

- Also:

$$r = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

where $\mathrm{cov}(x, y)$ is the *covariance* of x and y, defined by:

$$\mathrm{cov}(x, y) = x\bar{y} - \bar{x}\bar{y}$$

mean(product)-product(means)

SD line

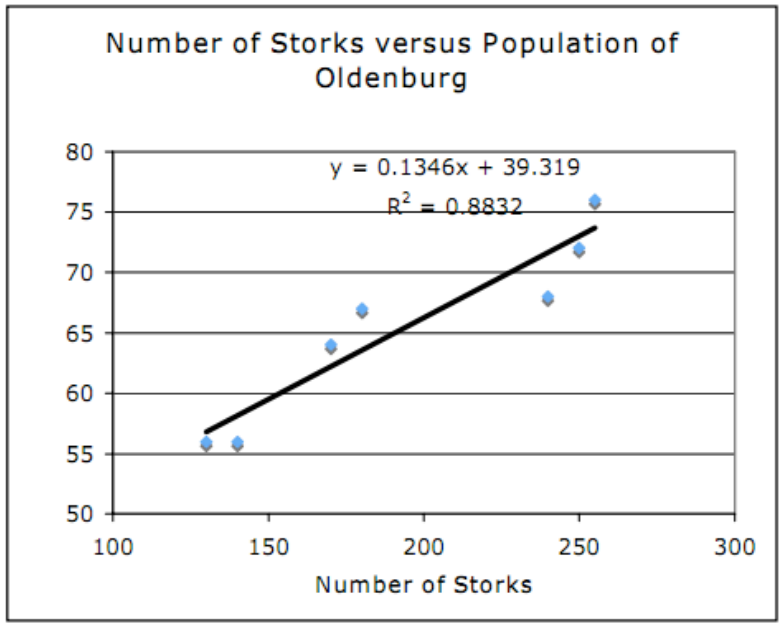
- The SD line passes through the *point of averages* $(\bar{x}, \bar{y})$
- Goes up 1 SD in y for each 1 SD in x, so plots the points $z_x = z_y$

$$y = \frac{s_y}{s_x}x + (\bar{y} - \frac{s_y}{s_x}\bar{x})$$

- SD line indicates drift of scatter diagram
- Slope too high, not good for prediction.
- Example: In the US, average male height is 69.2 inches, with an SD of 6.57 inches. Average weight is about 190 lb, with an SD of about 59.1 lb.
  - If a person is 1 SD above average height, then they are about 76 inches tall (6 feet 4 inches, 193 cm).
  - To lie on the SD line, they must be about 249 lb (113 kg), which is unusual.
  - Since height and weight are not *perfectly* correlated, SD line is not a good predictor.
- Note, if $r < 0$, SD line (and the scatter diagram) slope *down*.

The correlation coefficient r

1. r measures the strength of the **linear** relationship between the variables.
2. Very affected by outliers.
3. $|r| \leq 1$
4. $\text{cor}(x, y) = \text{cor}(y, x)$
5. r has no units, and so is unaffected by changes of scale.
6. correlation **does not imply** causation:

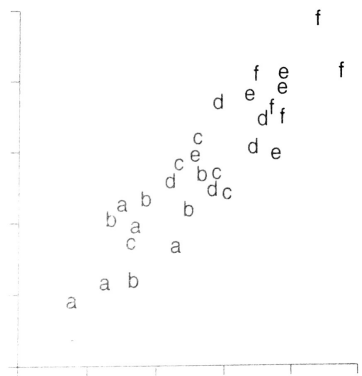


r

Example: B1, page 145:

xercise Set B

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?



Note that each group's correlation is not so high, but taken all together the correlation is much higher. This is called **reverse attenuation** (weakening).
When you restrict one of the variables correlation decreases, this is attenuation.

Ecological correlation

- This is correlation based on rates or averages, that tend to overstate r. The units are not individuals, but groups.
- *Ecological fallacy*: Assuming that relationships that hold for groups will also hold for individuals.
  1. If countries with more Protestants tend to have higher suicide rates, then Protestants must be more likely to commit suicide. (Inference from a 19th century study in Europe.)
  2. In a 1950 study, the relation between nativity and literacy was studied in each of the 48 states (of 1930 US). Two numbers were compared:
     i. The percent of foreign-born residents
     ii. The percent of literate residents

     r was found to be 0.53, indicating perhaps that the foreign-born tended to be more literate. But then, on further investigation, r for the individuals was -0.11! How is this?

Categorical variables

- So far, we have discussed scatterplots, and the correlation coefficient, which apply to quantitative data.
- What about categorical variables?
- Example: Are pesticides less often present in organic foods?
- Two categorical variables:
  - Pesticide status (present, not present)
  - Food status (organic, conventional)

	Pesticide Status		
Food Status	Present	Not Present	Total
Organic	29	98	127
Conventional	19485	7086	26571
Total	19514	7184	26698

Categorical variables

- Response? Explanatory?
- Overall proportion of foods with pesticide =
- Organic foods with pesticide =
- Conventional foods with pesticides =
- **Contingency tables** : 2 categorical variables.
- (row, column) : **cell**
- **cross-tabulation**: taking frequencies for cells of contingency tables.

Contingency tables

- We can find proportions of pesticide-containing foods for each type. These are called **conditional proportions** (the proportions are formed *conditional upon* each food type).
- Table of Conditional Proportions on Pesticide Status, for two food types:

	Pesticide	No pesticide	Total
Organic	0.23	0.77	127
Conventional	0.73	0.27	26571

- Note that the rows add up to 1. They must, these are proportions of a whole.
- If we just looked at row and column totals, regardless of food type, and wrote those in the **margins**, we would have **Marginal Proportions**.
- Usually, if there is no association, proportions would be about the same across the table. We will come back to this idea when we are doing Chi-squared tests.

- Example: A study in the UK in 1972-1974 asked 1314 British women if they smoked. 20 years later, there was a follow up survey checking if they were still alive. 24% of the smokers had died, and 31% of the non-smokers. Can we conclude that smoking is, in fact, good for your health?
- Correlation does not imply causation, and neither does association (of categorical variables) imply causation.

	Survival Status		
Smoker Status	Dead	Alive	Total
Yes	139	443	582
No	230	502	732
Total	369	945	1314

- Suppose we introduce the variable age, and look at survival status at different levels of age:

	18 – 34 yrs		35 – 54 yrs		55 – 64 yrs		65+ yrs	
Smoker Status	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive
Yes	5	174	41	198	51	64	42	7
No	6	213	19	180	40	81	165	28

- For each group, the percentage of smokers that died is higher.
- *Simpson's Paradox:* Direction of association between 2 variables can change after including a third variable and analyzing the data at separate levels of that variable.