Correlation and Covariance

James H. Steiger

Goals for Today

- Introduce the statistical concepts of
  - Covariance
  - Correlation
- Investigate invariance properties
- Develop computational formulas

Covariance

- So far, we have been analyzing summary statistics that describe aspects of a single list of numbers
- Frequently, however, we are interested in how variables behave together

Smoking and Lung Capacity

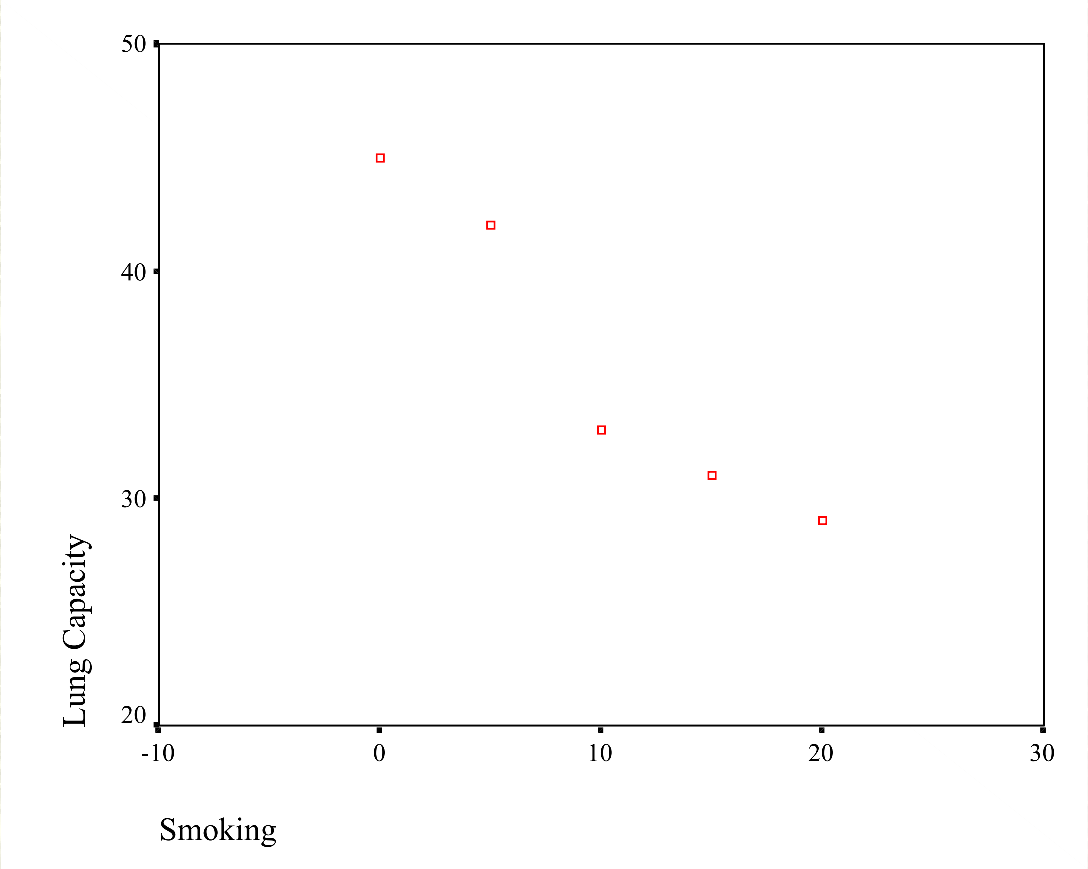
- Suppose, for example, we wanted to investigate the relationship between cigarette smoking and lung capacity
- We might ask a group of people about their smoking habits, and measure their lung capacities

Smoking and Lung Capacity

Cigarettes (*X*)	Lung Capacity (*Y*)
0	45
5	42
10	33
15	31
20	29

Smoking and Lung Capacity

- With SPSS, we can easily enter these data and produce a *scatterplot*.



Smoking and Lung Capacity

- We can see easily from the graph that as smoking goes up, lung capacity tends to go down.
- The two variables *covary* in opposite directions.
- We now examine two statistics, *covariance* and *correlation,* for quantifying how variables covary.

Covariance

- When two variables *covary* in opposite directions, as smoking and lung capacity do, values tend to be on opposite sides of the group mean. That is, when smoking is above its group mean, lung capacity tends to be below its group mean.
- Consequently, by averaging the product of deviation scores, we can obtain a measure of how the variables vary together.

The Sample Covariance

- Instead of averaging by dividing by N, we divide by $N-1$. The resulting formula is

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}_{\bullet}\right)\left(Y_i - \bar{Y}_{\bullet}\right)$$

Calculating Covariance

Cigarettes (X)	dX	$dXdY$	dY	Lung Capacity (Y)
0	–10	–90	+9	45
5	–5	–30	+6	42
10	0	0	–3	33
15	+5	–25	–5	31
20	+10	–70	–7	29

–215

Calculating Covariance

- So we obtain

$$S_{xy} = \frac{1}{4}(-215) = -53.75$$

Invariance Properties of Covariance

- The covariance is invariant under listwise addition, but *not* under listwise multiplication. Hence, it is vulnerable to changes in standard deviation of the variables, and is not *scale-invariant*.

Invariance Properties of Covariance

If $L_i = aX_i + b$, then

$$dl_i = adx_i$$

Let $L_i = aX_i + b,\ M_i = cY_i + d$

Then $S_{LM} = \dfrac{1}{N-1}\sum_{i=1}^{N} dl_i dm_i$

$$= \frac{1}{N-1}\sum_{i=1}^{N} adx_i\, cdy_i = ac\frac{1}{N-1}\sum_{i=1}^{N} dx_i\, dy_i = acS_{xy}$$

Invariance Properties of Covariance

- Multiplicative constants come straight through in the covariance, so covariance is difficult to interpret – it incorporates information about the scale of the variables.

The (Pearson) Correlation Coefficient

- Like covariance, but uses *Z*-scores instead of deviations scores. Hence, it is invariant under linear transformation of the raw scores.

$$r_{xy} = \frac{1}{N-1}\sum_{i=1}^{N} zx_i zy_i$$

Alternative Formula for the Correlation Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Computational Formulas -- Covariance

- There is a computational formula for covariance similar to the one for variance. Indeed, the latter is a special case of the former, since variance of a variable is "its covariance with itself."

$$s_{xy} = \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i Y_i - \frac{\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N} \right)$$

Computational Formula for Correlation

- By substituting and rearranging, you obtain a substantial (and not very transparent) formula for r_{xy}

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

Computing a correlation

Cigarettes (X)	X^2	XY	Y^2	Lung Capacity (Y)
0	0	0	2025	45
5	25	210	1764	42
10	100	330	1089	33
15	225	465	961	31
20	400	580	841	29
50	750	1585	6680	180

Computing a Correlation

$$r_{xy} = \frac{(5)(1585) - (50)(180)}{\sqrt{\left[(5)(750) - 50^2\right]\left[(5)(6680) - 180^2\right]}}$$

$$= \frac{7925 - 9000}{\sqrt{(3750 - 2500)(33400 - 32400)}}$$

$$= \frac{-1075}{\sqrt{(1250)(1000)}} = -.9615$$