
Correlation and Covariance

R. F. Riesenfeld

(Based on web slides by
James H. Steiger)

Goals

- ⇒ Introduce concepts of
 - Covariance
 - Correlation

- ⇒ Develop computational formulas

Covariance

- ⇒ Variables may change in relation to each other
- ⇒ *Covariance* measures how much the movement in one variable predicts the movement in a corresponding variable

Smoking and Lung Capacity

- ⇒ Example: investigate relationship between *cigarette smoking* and *lung capacity*
- ⇒ Data: sample group response data on smoking habits, *and* measured lung capacities, respectively

Smoking v Lung Capacity Data

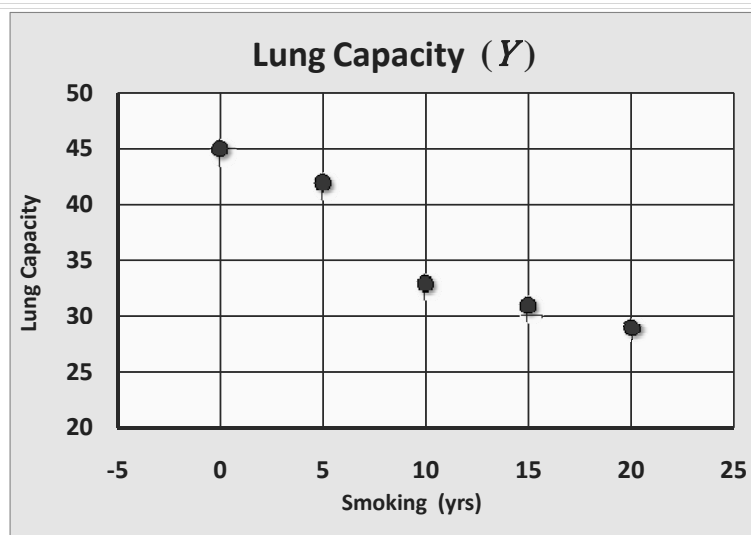
N	Cigarettes (X)	Lung Capacity (Y)
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

R F Riesenfeld Sp 2010

CS5961 Comp Stat

5

Smoking and Lung Capacity



6

Smoking v Lung Capacity

- ⇒ Observe that as smoking exposure goes up, corresponding lung capacity goes down
- ⇒ Variables *covary* inversely
- ⇒ *Covariance* and *Correlation* quantify relationship

Covariance

- ⇒ Variables that *covary* inversely, like smoking and lung capacity, tend to appear on opposite sides of the group means
 - When smoking is above its group mean, lung capacity tends to be below its group mean.
- ⇒ Average *product of deviation* measures extent to which variables covary, the degree of linkage between them

The Sample Covariance

- ⇒ Similar to variance, for theoretical reasons, average is typically computed using $(N-1)$, not N . Thus,

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

Calculating Covariance

Cigs (X)	Lung Cap (Y)
0	45
5	42
10	33
15	31
20	29
$\bar{X} = 10$	$\bar{Y} = 36$

Calculating Covariance

Cigs (X)	$(X - \bar{X})$	$(X - \bar{X})(Y - \bar{Y})$	$(Y - \bar{Y})$	Cap (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
		$\Sigma = -215$		

R F Riesenfeld Sp 2010

CS5961 Comp Stat

11

Covariance Calculation (2)

Evaluation yields,

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

R F Riesenfeld Sp 2010

CS5961 Comp Stat

12

Covariance under Affine Transformation

Let $L_i = aX_i + b$ and $M_i = cY_i + d$. Then,

$$(\Delta l)_i = a(\Delta x)_i, \quad (\Delta m)_i = c(\Delta y)_i,$$

$$\text{where, } (\Delta u)_i \equiv u_i - \bar{u}.$$

Evaluating, in turn, gives,

$$S_{LM} = \frac{1}{N-1} \sum_{i=1}^N (\Delta l)_i (\Delta m)_i$$

Covariance under Affine Transf (2)

Evaluating further,

$$S_{LM} = \frac{1}{N-1} \sum_{i=1}^N (\Delta l)_i (\Delta m)_i$$

$$= \frac{1}{N-1} \sum_{i=1}^N a(\Delta x)_i c(\Delta y)_i$$

$$= ac \frac{1}{N-1} \sum_{i=1}^N (\Delta x)_i (\Delta y)_i$$

$$\therefore S_{LM} = acS_{xy}$$

(Pearson) Correlation Coefficient r_{xy}

⇒ Like covariance, but uses Z-values instead of deviations. Hence, invariant under linear transformation of the raw data.

$$r_{xy} = \frac{1}{N-1} \sum_{i=1}^N z_{x_i} z_{y_i}$$

Alternative (common) Expression

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

Computational Formula 1

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

R F Riesenfeld Sp 2010

CS5961 Comp Stat

17

Computational Formula 2

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

R F Riesenfeld Sp 2010

CS5961 Comp Stat

18

Table for Calculating r_{xy}

Cigs (X)	X^2	XY	Y^2	Cap (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	
$\Sigma =$	50	750	1585	6680	180

R F Riesenfeld Sp 2010

CS5961 Comp Stat

19

Computing r_{xy} from Table

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

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

R F Riesenfeld Sp 2010

CS5961 Comp Stat

20

Computing Correlation

$$r_{xy} = \frac{-1075}{\sqrt{(1250)(1000)}}$$

$$r_{xy} = -0.9615$$

$r_{xy} = -0.96$ Conclusion

- ⇒ $r_{xy} = -0.96$ implies almost certainty smoker will have diminish lung capacity
- ⇒ Greater smoking exposure implies greater likelihood of lung damage

End
Covariance & Correlation
Notes