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

<table>
<thead>
<tr>
<th>$N$</th>
<th>Cigarettes ($X$)</th>
<th>Lung Capacity ($Y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>42</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>33</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>29</td>
</tr>
</tbody>
</table>

Smoking and Lung Capacity

Lung Capacity ($Y$)

Smoking (yrs)
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

<table>
<thead>
<tr>
<th>Cigs ((X))</th>
<th>Lung Cap ((Y))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>5</td>
<td>42</td>
</tr>
<tr>
<td>10</td>
<td>33</td>
</tr>
<tr>
<td>15</td>
<td>31</td>
</tr>
<tr>
<td>20</td>
<td>29</td>
</tr>
</tbody>
</table>

\(\bar{X} = 10\) \hspace{1cm} \(\bar{Y} = 36\)
### Calculating Covariance

<table>
<thead>
<tr>
<th>Cigs ((X))</th>
<th>((X - \overline{X}))</th>
<th>((X - \overline{X})(Y - \overline{Y}))</th>
<th>((Y - \overline{Y}))</th>
<th>Cap ((Y))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-10</td>
<td>-90</td>
<td>9</td>
<td>45</td>
</tr>
<tr>
<td>5</td>
<td>-5</td>
<td>-30</td>
<td>6</td>
<td>42</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>33</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>-25</td>
<td>-5</td>
<td>31</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>-70</td>
<td>-7</td>
<td>29</td>
</tr>
</tbody>
</table>

\[ \sum = -215 \]

### Covariance Calculation \((2)\)

Evaluation yields,

\[
S_{xy} = \frac{1}{4}(-215) = -53.75
\]
Covariance under Affine Transformation

Let \( L_i = aX_i + b \) and \( M_i = cY_i + d \). Then,

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

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,

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

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

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

\[
\therefore S_{LM} = ac S_{xy}
\]
(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

\[ S_{xy} = \frac{1}{N-1} \left( \sum_{i=1}^{N} X_i Y_i - \frac{\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N} \right) \]

Computational Formula 2

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

<table>
<thead>
<tr>
<th>Cigs ($X$)</th>
<th>$X^2$</th>
<th>$XY$</th>
<th>$Y^2$</th>
<th>Cap ($Y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2025</td>
<td>45</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>210</td>
<td>1764</td>
<td>42</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>330</td>
<td>1089</td>
<td>33</td>
</tr>
<tr>
<td>15</td>
<td>225</td>
<td>465</td>
<td>961</td>
<td>31</td>
</tr>
<tr>
<td>20</td>
<td>400</td>
<td>580</td>
<td>841</td>
<td>29</td>
</tr>
<tr>
<td>∑=</td>
<td>50</td>
<td>750</td>
<td>1585</td>
<td>6680</td>
</tr>
</tbody>
</table>

Computing $r_{xy}$ from Table

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

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

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

\[ r_{xy} = -0.9615 \]

\[ r_{xy} = -0.96 \quad \text{Conclusion} \]

\[ r_{xy} = -0.96 \text{ implies almost certainty} \]
\[ \text{smoker will have diminish lung capacity} \]

\[ \Rightarrow \text{Greater smoking exposure implies greater} \]
\[ \text{likelihood of lung damage} \]
End

Covariance & Correlation

Notes