Correlation

Unit 14

Quantifying the relation between two quantitative variables
Illustrative Data *(bicycle.sav)*

p. 14.1

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name1</td>
<td>32</td>
<td>384</td>
</tr>
<tr>
<td>Name2</td>
<td>10</td>
<td>354</td>
</tr>
<tr>
<td>Name3</td>
<td>5</td>
<td>233</td>
</tr>
<tr>
<td>Name4</td>
<td>11</td>
<td>225</td>
</tr>
<tr>
<td>Name5</td>
<td>21</td>
<td>514</td>
</tr>
<tr>
<td>Name6</td>
<td>81</td>
<td>316</td>
</tr>
<tr>
<td>Name7</td>
<td>13</td>
<td>218</td>
</tr>
<tr>
<td>Name8</td>
<td>50</td>
<td>454</td>
</tr>
<tr>
<td>Name9</td>
<td>10</td>
<td>555</td>
</tr>
<tr>
<td>Name10</td>
<td>5</td>
<td>215</td>
</tr>
<tr>
<td>Name11</td>
<td>11</td>
<td>320</td>
</tr>
<tr>
<td>Name12</td>
<td>20</td>
<td>331</td>
</tr>
</tbody>
</table>

*Neighborhood* *(0.016676136 + 0.015991116)* *(0.06098806 + 0.06099806)*
Scatterplot

Outliers can be important!
(Here we have a low SES school with a good outcome.)
Scatterplot (w/ outlier removed)

But, for purpose of quantifying linear relation, outliers has been removed.
Direction of Correlation

(A) Positive

(B) Negative

(C) None

(D) None
Strength of Correlation

(A) Strong Positive

(B) Weak Positive

(C) Strong Negative

(D) Weak Negative

Cannot determine strength visually (susceptible to axis-scaling)
Pearson’s Correlation Coefficient ($r$)

$p. 14.4$

$r$ varies from $-1$ to $1$
- $+1 = \text{perfect positive correlation (all points on upward line)}$
- $0 = \text{no correlation (flat scatter cloud)}$
- $-1 = \text{perfect negative correlation (all points on downward line)}$
Strong and weak correlations

The closer $|r|$ is to 1, the “stronger” the correlation

(A) Strong Positive Correlation      (B) Weak Positive Correlation

(C) Strong Negative Correlation      (D) Weak Negative Correlation
“Rules-of-thumb”
(offered with hesitation)

\[0.0 \leq |r| < 0.3 \rightarrow \text{weak correlation}\]
\[0.3 \leq |r| < 0.7 \rightarrow \text{moderate correlation}\]
\[0.7 \leq |r| < 1.0 \rightarrow \text{strong correlation}\]
\[|r| = 1.0 \rightarrow \text{perfect correlation}\]

\[|r|\] denotes the absolute value of \(r\)
Calculating the Correlation Coefficient

Three different sums of squares

- Sum of Squares of X variable ($SS_{XX}$) = $\sum (x_i - xbar)^2$
- Sum of Squares of Y variable ($SS_{YY}$) = $\sum (y_i - ybar)^2$
- Sum of Cross-product ($SS_{XY}$) = $\sum (x_i - xbar)(y_i - ybar)$

- $r = SS_{XY} / \sqrt{[(SS_{XX})(SS_{YY})]}$

Illustrative example (*bicycle.sav*)

- $SS_{XX} = 7855.67$
- $SS_{YY} = 3159.68$
- $SS_{XY} = -4231.1333$
- $r = -4231.1333/\sqrt{[(7855.67)(3159.68)]} = -0.85$

Interpretation: strong negative correlation
Coefficient of determination ($r^2$)

- $r^2 = \text{coefficient of determination}$
- Interpretation = variation of Y “explained” by X
  - Caveat: the “explanation” is merely mathematical and is not necessarily causal
  - Correlation ≠ causation!
- Illustrative example (bicycle.sav)
  - $r^2 = -0.85^2 = 0.72$
  - 72% of variability in helmet use ”explained” by SES
Let $\rho$ (rho) represent the correlation coefficient parameter
- Recall, $r =$ observed (sample) correlation
- $\rho$ can be 0 while $r$ is non-0 (figure below)
$H_0: \rho = 0$ versus $H_1: \rho \neq 0$

with df = $n - 2$

$$z_{6.1} = \sqrt{\frac{n-3}{n-1}} \cdot \frac{z_{6.1}}{e_{1.5}}$$

- Using these formulas (`bicycle.sav`)
  - $se_r = \sqrt{[1 - 0.849^2] / (12 - 2)} = 0.167$
  - $t_{\text{stat}} = (0.849 / 0.167) = -5.08$
  - $df = 12 - 2 = 10$
  - $p = .00048$ (using StaTable)
  - Conclude: $\rho$ is negative ($r$ is significant)
Assumptions for testing

- Validity assumptions (of course)
- Distributional assumptions
  - Linearity
  - Independence in sampling
  - Bivariate normality (figure)
Correlation $\neq$ Causation

- All statistical relations mean little without a deeper look into causation
- Causal inference is a separate objective (somewhat beyond scope of this unit)
- So what does really matter?
  - Intellectual content
  - Organization of data
  - Reliance on tests of experience
Mean ± se