The Interpretation and Utility of Three Cohesion Metrics for Object-Oriented Design

Steve Counsell, Stephen Swift, and Jason Crampton

*ACM TOSEM, April 2006*

2006. 7. 26
Taehoon Song
Contents

- Introduction
- Motivation and related work
- Preliminaries
- Cohesion metrics
- Empirical results
- Conclusion and further work
- Discussion
Introduction (1/2)

- Definition of cohesion
  - Is regard a **class as cohesive if the methods of the class use the same set of parameter types** [1]

- Metrics for measuring cohesion
  - Cohesion among methods in a class metric (CAMC) [1]
  - Normalised hamming distance metric (NHD) [2]
  - Scaled NHD metric

Purpose of this article

- Rigorous mathematical analysis
  - Determine whether these metrics have any qualitative meaning given the definition of cohesion above
  - Determine what values of these metrics should be considered to represent a cohesive class
Motivation and related work (1/2)

- Mathematical comparison of the properties of cohesion metrics
  - Is an **under researched area**
- Identifying common failings or properties of cohesion metrics
  - Informs our understanding of OO system, OO language and their different traits

Examination and scrutiny of **current** cohesion metrics
Motivation and related work (2/2)

- Object-oriented paradigm
  - Notion of class cohesion has superseded that of module of cohesion
    - Lack of cohesion of methods metric (LCOM) [3]
      - On the assumption that a class is cohesive if the same instance variables appear in most or all of the methods in a class
        - Values produced by the metric are difficult to interpret and give little insight into the nature of the class
        - The metric is an implementation metric
          - Measure of cohesion is required earlier in the development process (at design time)

Preliminaries (1/4)

- Three systems
  - Represent a variety of different application domains
    - Edge (graph editor: 30.8 KNCSL, 80 classes)
    - Rocket (compiler: 32.4 KNCSL, 322 classes)
    - Et++ (user interface framework: 56.3 KNCSL, 508 classes)
  
- Using classes randomly chosen from these systems
Preliminaries (2/4)

- Notion of an entity relational system (ERS)
  - Provides a mapping from the real world attribute of the entity being measured to values in the empirical world
  - Assigns a measure of the similarity between the parameter types of the methods for each class
    - A class X is more cohesive than class Y if this function returns a higher value for X when there is greater sharing of parameters between the methods of a class
  - Distinguish between the cohesiveness of n classes using the same metric
Preliminaries (3/4)

- **Notation**
  - $(i, j)$th entry of a matrix $M : m_{ij}$
  - Given class : $C$
    - Methods : $k$ (*row vector*)
    - Parameter type list : $L$ (*column vector*)
      - Length of $L : l$
  - Parameter occurrence matrix $O$ (1≤$i$≤$k$, 1≤$j$≤$l$)

$$O_{ij} = \begin{cases} 
1 & \text{if the } j\text{th data type occurs as a parameter in the } i\text{th method} \\
0 & \text{otherwise}
\end{cases}$$
Preliminaries (4/4)

- Notation (Cont’d)

```
Alert(AlertType, byte, *text=0, Bitmap *bm=0);
~Alert();
VObject *DoCreateDialog();
int Show(char *fmt);
int ShowV(char *fmt, va_list ap);
class Menu *GetMenu();
void InspectorId(char *buf, int sz);
```

(a) Methods

- **Binary $k \times l$ matrix**
  - $i$th row: parameter occurrence vector (for method $i$)
  - Indicates the presence of data types in the $i$th method

```
Alert
1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1
0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 1 0
```

(b) Parameter occurrence matrix

```
$R_i = \sum_{j=1}^{l} O_{ij}$, $C_j = \sum_{i=1}^{k} O_{ij}$, $\sigma = \sum_{i=1}^{k} \sum_{j=1}^{l} O_{ij}$
```
Cohesion metrics (1/6)

- CAMC
  - Can be evaluated at design time ($\leftrightarrow LCOM$)
  - Is the average of the entries in the parameter occurrence matrix
  - Formulation

$$\text{CAMC}(C) = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} o_{ij} = \frac{\sigma}{kl}$$

Cohesion indicator: 0.35

$$\text{CAMC}(\text{Alert}) = \frac{1}{7*6} (3+0+0+1+2+0+2) = \frac{8}{42} \approx 0.19$$
Cohesion metrics (2/6)

- CAMC (Cont’d)
  - CAMCs
    - Includes the “self” parameter type in the parameter occurrence matrix
      - Append a column of 1s to the parameter occurrence matrix forming the \((l+1)\)th column

\[
\text{CAMCs}(C) = \frac{\sigma + k}{K(l+1)}
\]

\[
\text{CAMC}(C) = \frac{\sigma}{kl}
\]

\[
\text{CAMCs(Alert)} = \frac{8+7}{7*(6+1)} = 15/49 \approx 0.31
\]
Cohesion metrics (3/6)

- **NHD**
  - Measures agreement between rows in a binary matrix
  - Alternative measure of the cohesion in the sense computed by the CAMC metric
  - Parameter agreement matrix $A$

\[
A = \begin{cases} 
.aij & \text{if the parameter agreement is between methods } i \text{ and } j \ (1 \leq j < i \leq k) \\
0 & \text{otherwise}
\end{cases}
\]
Cohesion metrics (4/6)

- NHD (Cont’d)
  - Parameter agreement matrix
    - Lower triangular square matrix of dimension $k-1$

(c) Formulation

\[
NHD(C) = \frac{1}{l(k^{\frac{1}{2}})} \sum_{j=1}^{k-1} \sum_{i=j+1}^{k} a_{ij} = \frac{2}{lk(k-1)} \sum_{j=1}^{k-1} \sum_{i=j+1}^{k} a_{ij}
\]

NHD(Alert) = \( \frac{2}{6*7*(7-1)} \times 84 = 168/252 \approx 0.67 \)

(a) Parameter occurrence matrix

(b) Parameter agreement matrix
Cohesion metrics (5/6)

- NHD (Cont’d)

- Can distinguish between different parameter occurrence matrices with the same number of 1s

  - $\text{NHD}_{\text{min}} \leq \text{NHD} \leq \text{NHD}_{\text{max}}$ (l ≤ σ ≤ kl)

  \[ \text{NHD}(C) = \frac{1}{l(k)} \sum_{j=1}^{k} \sum_{i=j+1}^{k} a_{ij} = \frac{2}{lk(k-1)} \sum_{j=1}^{k} \sum_{i=j+1}^{k} a_{ij} \]

  \[ \text{NHD}_{\text{min}} = 1 - \frac{q(d+1)(k-d-1)+l(q)d(k-d)}{lk(k-1)} \]

  \[ \text{NHD}_{\text{max}} = 1 - \frac{r(1+d)(k-r-1)+(l+c)d(k-d)}{lk(k-1)} \]

  \[ \sigma = 8 \]

  \[ k = 7 \]

  \[ l = 6 \]

<table>
<thead>
<tr>
<th>2</th>
<th>0</th>
<th>4</th>
<th>0</th>
<th>4</th>
<th>0</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Parameter agreement matrix

\[ \begin{array}{ccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array} \]

\[ \begin{array}{ccccccc}
1 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\end{array} \]

\[ m_1 = \begin{array}{ccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array} \quad m_2 = \begin{array}{ccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array} \]

\[ m_3 = \begin{array}{ccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array} \quad m_4 = \begin{array}{ccccccc}
1 & 6 & 6 & 6 & 6 & 6 & 6 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
\end{array} \]

\[ m_5 = \begin{array}{ccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array} \]

\[ m_6 = \begin{array}{ccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array} \]

\[ d = \begin{array}{ccccccc}
1 & 1 & \sigma \sigma \sigma \sigma \sigma \sigma \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{array} \quad \hat{r} = \begin{array}{ccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array} \]

Parameter agreement matrix

\[ \begin{array}{ccccccc}
1 & 6 & 6 & 6 & 6 & 6 & 6 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
0 & 5 & 5 & 5 & 5 & 5 & 5 \\
\end{array} \]

82 < 84
Cohesion metrics (6/6)

- **Scaled NHD**
  - Uses of the fact that both ends of the range of values for NHD
  - Represents how close the NHD metric is to the maximum value of NHD compared to the minimum value

\[
SNHD = \begin{cases} 
0 & \text{if } NHD_{\text{min}} = NHD_{\text{max}} \text{ and } \sigma < kl \\
1 & \text{if } \sigma = kl \\
\frac{NHD - NHD_{\text{min}}}{NHD_{\text{max}} - NHD_{\text{min}}} - 1 & \text{otherwise}
\end{cases}
\]
Empirical results (1/2)

Evaluation of cohesion metrics

<table>
<thead>
<tr>
<th>System</th>
<th>Class</th>
<th>$k$</th>
<th>$l$</th>
<th>CAMC$_a$</th>
<th>CAMC</th>
<th>NHD$_a$</th>
<th>NHD</th>
<th>SNHD$_a$</th>
<th>SNHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Et++</td>
<td>Alert</td>
<td>7</td>
<td>6</td>
<td>0.306</td>
<td>0.190</td>
<td>0.714</td>
<td>0.667</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>ApplDialog</td>
<td>4</td>
<td>3</td>
<td>0.438</td>
<td>0.250</td>
<td>0.625</td>
<td>0.500</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>BagItem</td>
<td>11</td>
<td>4</td>
<td>0.309</td>
<td>0.136</td>
<td>0.804</td>
<td>0.755</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Dialog</td>
<td>15</td>
<td>6</td>
<td>0.248</td>
<td>0.122</td>
<td>0.810</td>
<td>0.778</td>
<td>0.830</td>
<td>-0.586</td>
</tr>
<tr>
<td></td>
<td>CycleItem</td>
<td>14</td>
<td>11</td>
<td>0.202</td>
<td>0.130</td>
<td>0.797</td>
<td>0.778</td>
<td>0.512</td>
<td>-0.451</td>
</tr>
<tr>
<td></td>
<td>BitMap</td>
<td>22</td>
<td>10</td>
<td>0.169</td>
<td>0.086</td>
<td>0.856</td>
<td>0.842</td>
<td>0.757</td>
<td>-0.555</td>
</tr>
<tr>
<td></td>
<td>Assoc</td>
<td>11</td>
<td>3</td>
<td>0.409</td>
<td>0.212</td>
<td>0.773</td>
<td>0.697</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Rocket</td>
<td>Arc</td>
<td>5</td>
<td>2</td>
<td>0.467</td>
<td>0.260</td>
<td>0.733</td>
<td>0.600</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>ArcList</td>
<td>9</td>
<td>3</td>
<td>0.389</td>
<td>0.185</td>
<td>0.764</td>
<td>0.685</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>CallGraph</td>
<td>11</td>
<td>4</td>
<td>0.273</td>
<td>0.091</td>
<td>0.855</td>
<td>0.818</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>DDGктTypeList</td>
<td>9</td>
<td>3</td>
<td>0.389</td>
<td>0.185</td>
<td>0.764</td>
<td>0.685</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>DDGGnodePtrList</td>
<td>10</td>
<td>3</td>
<td>0.475</td>
<td>0.300</td>
<td>0.661</td>
<td>0.548</td>
<td>0.227</td>
<td>-0.835</td>
</tr>
<tr>
<td></td>
<td>DataType</td>
<td>20</td>
<td>5</td>
<td>0.225</td>
<td>0.070</td>
<td>0.887</td>
<td>0.864</td>
<td>0.989</td>
<td>-1.000</td>
</tr>
<tr>
<td></td>
<td>DeclaratorPtrList</td>
<td>11</td>
<td>3</td>
<td>0.477</td>
<td>0.303</td>
<td>0.604</td>
<td>0.552</td>
<td>0.177</td>
<td>-0.875</td>
</tr>
<tr>
<td>Edge</td>
<td>null_dummy</td>
<td>7</td>
<td>4</td>
<td>0.343</td>
<td>0.179</td>
<td>0.733</td>
<td>0.667</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>constr_descriptor</td>
<td>8</td>
<td>9</td>
<td>0.225</td>
<td>0.139</td>
<td>0.757</td>
<td>0.730</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>constr_queue</td>
<td>8</td>
<td>8</td>
<td>0.292</td>
<td>0.203</td>
<td>0.687</td>
<td>0.647</td>
<td>0.344</td>
<td>-0.807</td>
</tr>
<tr>
<td></td>
<td>constr_manager</td>
<td>17</td>
<td>8</td>
<td>0.229</td>
<td>0.132</td>
<td>0.827</td>
<td>0.805</td>
<td>0.773</td>
<td>0.206</td>
</tr>
<tr>
<td></td>
<td>gne_default</td>
<td>14</td>
<td>3</td>
<td>0.393</td>
<td>0.190</td>
<td>0.775</td>
<td>0.700</td>
<td>0.868</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>elist</td>
<td>10</td>
<td>2</td>
<td>0.533</td>
<td>0.300</td>
<td>0.704</td>
<td>0.556</td>
<td>0.521</td>
<td>-0.490</td>
</tr>
<tr>
<td></td>
<td>intersect</td>
<td>21</td>
<td>4</td>
<td>0.314</td>
<td>0.143</td>
<td>0.800</td>
<td>0.750</td>
<td>0.758</td>
<td>-0.750</td>
</tr>
</tbody>
</table>
Empirical results (2/2)

- Cross comparison of the three metrics
  - Correlations between the three cohesion metrics

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Pearson’s</th>
<th>Spearman’s</th>
<th>Kendall’s</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCOM versus SNHD</td>
<td>-0.458</td>
<td>-0.425</td>
<td>-0.337</td>
<td>-0.407</td>
</tr>
<tr>
<td>LCOM versus CAMC</td>
<td>-0.540</td>
<td>-0.0239</td>
<td>-0.156</td>
<td>-0.312</td>
</tr>
</tbody>
</table>

Table shows evidence of correlation between design-time metrics (such as CAMC and SNHD) and code-based metrics (such as LCOM)

- SNHD metric produces higher results than CAMC
Conclusion

Further work

- Undertake a more formal and extensive analysis of the SNHD metric
- Conduct more extensive tests on whole systems
- Establish significant values of the SNHD metric
Discussion

- Need to make a best definition of cohesion
  - What cohesion means in object-oriented software in light of the matrices $O_{\text{max}}$ and $O_{\text{min}}$
- Need to adjust more system
  - Exclude prior knowledge (Edge, rocket, et++)
  - Use more various number of classes
    - Threat to scalability of the results
  - Choose classes in more OO application types
- Need to interplay between cohesion and OO coupling
Background

- **Original definition of the metric in OO sense**
  - Calculates cohesion according to use of class attributes in the methods of a class
  - Is based on the principle that an instance variable occurring in many methods of a class
    - Causes that class to be more cohesive than one where the same variable is used in very few methods of the class

- **High cohesion**
  - Reflects using of development technique known to produce robust and maintainable code
Procedural programming viewpoint

- Modules were the key elements by which cohesion was measured
  - Inter-module metrics (Stevens et al. [1974])
  - Seven point ordinal scale for component cohesion (Yourdon and Constantine [1979])
    - Functional cohesion
      - Module perform a single well-defined function
    - Coincidental cohesion
      - Module perform more than one function, and that those functions were unrelated
Structured paradigm

- Informal definition of cohesion (Lakhotia [1993])
  - Was built on the basis of sound programmer practice and experience
  - Was underpinned work on measuring module cohesion
CAMC

- Interpreting
  - Can’t distinguish between the cohesion of different matrices with the same value of $\sigma$
  - Has fundamentally flawed about using 0.35 as a threshold for an indicator
  - Is likely to find smaller classes more cohesive, irrespective of their actual properties

<Minimum values of CAMC and CAMCs >
Cohesion metrics (2/2)

- Hamming distance (HD) metric
  - Provides a measure of disagreement between rows in a binary matrix informally (Counsell et al. [2001])

- NHD
  - Interpretation
    - Suggest that a class for which the NHD metric is more than 0.5 should be considered cohesive (Counsell et al. [2002])
    - Must reconsider carefully what we mean by cohesion
      - It is questionable whether this is satisfactory behavior for a cohesion metric, as small classes are generally regarded as being more cohesive than large ones
Correlation (correlation coefficient)

- Indicates the strength and direction of a linear relationship between two random variables
- Random variable
  - is a mathematical function that maps outcomes of random experiments to numbers
Pairwise independence

- Collection of random variables is a set of random variables any two of which are independent
  - Suppose X, Y and Z have the following joint probability distribution
    \[
    (X, Y, Z) = \begin{cases} 
    (0, 0, 0) & \text{with probability 1/4}, \\
    (0, 1, 1) & \text{with probability 1/4}, \\
    (1, 0, 1) & \text{with probability 1/4}, \\
    (1, 1, 0) & \text{with probability 1/4}.
    \end{cases}
    \]
  - X-Y, X-Z, Y-Z are independent
  - X, Y, Z are not independent
    - Mod 2 sum of the other two is completely determined by other two
Correlation (3/5)

- Pearson’s correlation coefficient
  - Is defined only if both of the standard deviations are finite and both of them are nonzero
    - Is a corollary of the Cauchy-Schwarz inequality that the correlation cannot exceed 1 in absolute value
  - Is 1 in the case of an increasing linear relationship, −1 in the case of a decreasing linear relationship
    - The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables.
  - If the variables are independent then the correlation is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.
Spearman’s correlation coefficient

- Is a non-parametric measure of correlation
- Assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables

\[
\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}
\]

\(D\) = the difference between the ranks of corresponding values of \(X\) and \(Y\), and
\(N\) = the number of pairs of values

*Spearman's rank correlation coefficient is equivalent to Pearson correlation on ranks*
Correlation (5/5)

Kendall’s correlation coefficient

\[ \tau = \frac{2P}{\frac{1}{2}n(n-1)} - 1 \]

Deals with measuring correspondence between two rankings, and assessing the significance of this correspondence between different rankings on the same set of items.

- If the agreement between the two rankings is perfect, the coefficient has value 1.
- If the disagreement between the two rankings is perfect, the coefficient has value -1.
- If the rankings are independent, the coefficient has value 0.

<table>
<thead>
<tr>
<th>Person</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank by height</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Rank by weight</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>