**Introduction to Statistics in Psychology**

PSY 201

Professor Greg Francis

Lecture 10

correlation

*How changes in one variable correspond to change in another variable.*

---

**CORRELATION**

two variables may be related

- SAT scores, GPA
- hours in therapy, self-esteem
- grade on homeworks, grade on exams
- number of risk factors, probability of getting AIDS
- height, points in basketball
- ...

how do we show the relationship?

scattergrams

---

**RELATIONSHIPS**

Identifying these types of relationships is one of the key issues in statistical analysis

Consider a 1999 study that reported a relationship between the use of nightlights in a child's room and the tendency of the child to need glasses

My daughter slept with a nightlight?

---

**COMPLICATIONS**

Clearly there is a relationship between using a nightlight and needing glasses

However, it’s not clear what the nature of the relationship involves

It *could* be that the extra light somehow influences the child’s eyes and causes the need for glasses

Or it could be that needing glasses will somehow co-occur with the use of a nightlight (e.g., children who need glasses will want a night light, or their parents will want a nightlight)

Finding a relationship is necessary for establishing causation, but it is not enough

---

**SPURIOUS CORRELATION**

Since so many variables get measured, it is easy to identify spurious correlations

Sometimes there is an explanation for the relationship:

(increased use of technology)
SPURIOUS CORRELATION
Since so many variables get measured, it is easy to identify spurious correlations.
Sometimes there is no explanation for the relationship.

POSITIVE CORRELATION
First, we need to understand how to quantify the existence of a relationship.
Increases in the value of one variable tend to occur with increases in the value of the other variable.
SAT scores and exam scores

NEGATIVE CORRELATION
Increases in the value of one variable tend to occur with decreases in the value of the other variable.
temperature and number of people with frostbite

PERFECT CORRELATIONS
perfect positive correlation

perfect negative correlation

NO CORRELATION
no correlation
balance of larger and smaller values

CORRELATION COEFFICIENT
quantitative measure of correlation bounded between

-1.0 & +1.0

- correlation coefficient of -1.0 indicates perfect negative correlation
- correlation coefficient of +1.0 indicates perfect positive correlation
- correlation coefficient of 0.0 indicates no correlation
- values in between give ordinal measures of relationship
PEARSON $r$

Pearson product-moment correlation coefficient

one correlation coefficient for quantitative data

(the most important one)

degree to which $X$ and $Y$ vary together

degree to which $X$ and $Y$ vary separately

several formulas

1. $z$-scores
2. Deviation scores
3. Raw scores
4. Covariance

all give the same result!

$z$ SCORES

Two steps

1. Convert raw scores into $z$ scores
2. Find the mean of cross-products

$$r_{xy} = \frac{\Sigma z_x z_y}{n - 1}$$

$z$ SCORES

what does this calculation do?
suppose you have two distributions that have a positive correlation

then a large value of $X$ will be above $\bar{X}$ and have a positive $z_x$ score

and a corresponding $Y$ will be above $\bar{Y}$ and have a positive $z_y$ score

Thus the cross-product $z_x z_y$ will be positive

while a small value of $X$ will be below $\bar{X}$ and have a negative $z_x$ score

and the corresponding $Y$ will be below $\bar{Y}$ and have a negative $z_y$ score

Thus $z_x z_y$ will again be negative

to find the average, sum all the products (negative numbers) we divide by $n - 1$

$$r_{xy} = \frac{\Sigma z_x z_y}{n - 1}$$

still a negative number!

PEARSON $r$

also a small value of $X$ will be below $\bar{X}$ and have a negative $z_x$ score

and the corresponding $Y$ will be below $\bar{Y}$ and have a negative $z_y$ score

Thus

$z_x z_y$

will again be positive

to find the average, sum all the products (positive numbers) we divide by $n - 1$

$$r_{xy} = \frac{\Sigma z_x z_y}{n - 1}$$

still a positive number!

PEARSON $r$

exactly the opposite is true for negatively correlated distributions

then a large value of $X$ will be above $\bar{X}$ and have a positive $z_x$ score

and a corresponding $Y$ will be below $\bar{Y}$ and have a negative $z_y$ score

Thus

$z_x z_y$

will be negative

to find the average, sum all the products (negative numbers) we divide by $n - 1$

$$r_{xy} = \frac{\Sigma z_x z_y}{n - 1}$$

still a negative number!
**DEVIAITION FORMULA**

It is awkward to convert to $z$ scores
we can get the same number with
deviation scores

\[
x = X - \bar{X}
\]

\[
y = Y - \bar{Y}
\]

deviation score formula

\[
r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}
\]

---

**RAW SCORE FORMULA**

It is awkward to calculate deviation scores

raw score formula

\[
r_{xy} = \frac{n \sum XY - \sum X \sum Y}{\sqrt{n \sum X^2 - (\sum X)^2} \sqrt{n \sum Y^2 - (\sum Y)^2}}
\]

---

**COVARIANCE FORMULA**

Covariance $= s_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{n - 1}$

average cross-product of deviation scores
(similar to variance)

Pearson $r$ turns out to be:

\[
r_{xy} = \frac{s_{xy}}{s_x s_y}
\]

where $s_x$ and $s_y$ are the standard deviations of their respective distributions

---

**EXAMPLE**

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>$z_x$</th>
<th>$z_y$</th>
<th>$z_x z_y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>505</td>
<td>68</td>
<td>0.63</td>
<td>1.64</td>
<td>1.03</td>
</tr>
<tr>
<td>520</td>
<td>55</td>
<td>-0.15</td>
<td>0.35</td>
<td>-0.05</td>
</tr>
<tr>
<td>715</td>
<td>65</td>
<td>1.88</td>
<td>1.34</td>
<td>2.52</td>
</tr>
<tr>
<td>485</td>
<td>42</td>
<td>-1.34</td>
<td>-0.94</td>
<td>1.26</td>
</tr>
<tr>
<td>690</td>
<td>64</td>
<td>1.51</td>
<td>1.24</td>
<td>1.87</td>
</tr>
<tr>
<td>490</td>
<td>45</td>
<td>-0.46</td>
<td>-0.64</td>
<td>0.29</td>
</tr>
<tr>
<td>565</td>
<td>56</td>
<td>0.32</td>
<td>0.45</td>
<td>0.14</td>
</tr>
<tr>
<td>580</td>
<td>59</td>
<td>0.48</td>
<td>0.74</td>
<td>0.36</td>
</tr>
<tr>
<td>615</td>
<td>56</td>
<td>0.64</td>
<td>0.47</td>
<td>0.38</td>
</tr>
<tr>
<td>435</td>
<td>42</td>
<td>-1.03</td>
<td>-0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>440</td>
<td>38</td>
<td>-0.97</td>
<td>-1.33</td>
<td>1.29</td>
</tr>
<tr>
<td>515</td>
<td>50</td>
<td>0.20</td>
<td>-0.15</td>
<td>0.03</td>
</tr>
<tr>
<td>380</td>
<td>37</td>
<td>-1.00</td>
<td>-1.43</td>
<td>2.29</td>
</tr>
<tr>
<td>519</td>
<td>45</td>
<td>-0.27</td>
<td>-0.94</td>
<td>0.24</td>
</tr>
<tr>
<td>565</td>
<td>54</td>
<td>0.32</td>
<td>0.45</td>
<td>0.14</td>
</tr>
</tbody>
</table>

\[
\sum z_x = 0.05 \quad \sum z_y = 0.46 \quad \sum z_x z_y = 12.67
\]

---

**EXAMPLE**

standard score formula

\[
r_{xy} = \frac{\sum z_x z_y}{n - 1} = 12.67 = 0.905
\]

\[
x\quad y\quad z_x\quad z_y
\]

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$z_x$</th>
<th>$z_y$</th>
<th>$z_x z_y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>505</td>
<td>68</td>
<td>0.63</td>
<td>1.64</td>
<td>1.03</td>
</tr>
<tr>
<td>520</td>
<td>55</td>
<td>-0.15</td>
<td>0.35</td>
<td>-0.05</td>
</tr>
<tr>
<td>715</td>
<td>65</td>
<td>1.88</td>
<td>1.34</td>
<td>2.52</td>
</tr>
<tr>
<td>485</td>
<td>42</td>
<td>-1.34</td>
<td>-0.94</td>
<td>1.26</td>
</tr>
<tr>
<td>690</td>
<td>64</td>
<td>1.51</td>
<td>1.24</td>
<td>1.87</td>
</tr>
<tr>
<td>490</td>
<td>45</td>
<td>-0.46</td>
<td>-0.64</td>
<td>0.29</td>
</tr>
<tr>
<td>565</td>
<td>56</td>
<td>0.32</td>
<td>0.45</td>
<td>0.14</td>
</tr>
<tr>
<td>580</td>
<td>59</td>
<td>0.48</td>
<td>0.74</td>
<td>0.36</td>
</tr>
<tr>
<td>615</td>
<td>56</td>
<td>0.64</td>
<td>0.47</td>
<td>0.38</td>
</tr>
<tr>
<td>435</td>
<td>42</td>
<td>-1.03</td>
<td>-0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>440</td>
<td>38</td>
<td>-0.97</td>
<td>-1.33</td>
<td>1.29</td>
</tr>
<tr>
<td>515</td>
<td>50</td>
<td>0.20</td>
<td>-0.15</td>
<td>0.03</td>
</tr>
<tr>
<td>380</td>
<td>37</td>
<td>-1.00</td>
<td>-1.43</td>
<td>2.29</td>
</tr>
<tr>
<td>519</td>
<td>45</td>
<td>-0.27</td>
<td>-0.94</td>
<td>0.24</td>
</tr>
<tr>
<td>565</td>
<td>54</td>
<td>0.32</td>
<td>0.45</td>
<td>0.14</td>
</tr>
</tbody>
</table>

\[
\sum z_x = 0.05 \quad \sum z_y = 0.46 \quad \sum z_x z_y = 12.67
\]

---

**EXAMPLE**

deviation score formula

\[
r_{xy} = \frac{\frac{\sum x y}{\sqrt{\sum x^2 \sqrt{\sum y^2}}}}{\sqrt{(130.000.00) \sqrt{1429.72}}} = 0.903
\]

\[
x\quad y\quad z_x\quad z_y
\]

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$z_x$</th>
<th>$z_y$</th>
<th>$z_x z_y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>505</td>
<td>68</td>
<td>0.63</td>
<td>1.64</td>
<td>1.03</td>
</tr>
<tr>
<td>520</td>
<td>55</td>
<td>-0.15</td>
<td>0.35</td>
<td>-0.05</td>
</tr>
<tr>
<td>715</td>
<td>65</td>
<td>1.88</td>
<td>1.34</td>
<td>2.52</td>
</tr>
<tr>
<td>485</td>
<td>42</td>
<td>-1.34</td>
<td>-0.94</td>
<td>1.26</td>
</tr>
<tr>
<td>690</td>
<td>64</td>
<td>1.51</td>
<td>1.24</td>
<td>1.87</td>
</tr>
<tr>
<td>490</td>
<td>45</td>
<td>-0.46</td>
<td>-0.64</td>
<td>0.29</td>
</tr>
<tr>
<td>565</td>
<td>56</td>
<td>0.32</td>
<td>0.45</td>
<td>0.14</td>
</tr>
<tr>
<td>580</td>
<td>59</td>
<td>0.48</td>
<td>0.74</td>
<td>0.36</td>
</tr>
<tr>
<td>615</td>
<td>56</td>
<td>0.64</td>
<td>0.47</td>
<td>0.38</td>
</tr>
<tr>
<td>435</td>
<td>42</td>
<td>-1.03</td>
<td>-0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>440</td>
<td>38</td>
<td>-0.97</td>
<td>-1.33</td>
<td>1.29</td>
</tr>
<tr>
<td>515</td>
<td>50</td>
<td>0.20</td>
<td>-0.15</td>
<td>0.03</td>
</tr>
<tr>
<td>380</td>
<td>37</td>
<td>-1.00</td>
<td>-1.43</td>
<td>2.29</td>
</tr>
<tr>
<td>519</td>
<td>45</td>
<td>-0.27</td>
<td>-0.94</td>
<td>0.24</td>
</tr>
<tr>
<td>565</td>
<td>54</td>
<td>0.32</td>
<td>0.45</td>
<td>0.14</td>
</tr>
</tbody>
</table>

\[
\sum z_x = 0.05 \quad \sum z_y = 0.46 \quad \sum z_x z_y = 12.67
\]
EXAMPLE

raw score formula

\[ r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{880.86}{96.53 \times 10.11} = 0.903 \]

where,

\[ s_{xy} = \frac{\sum xy}{n - 1} = \frac{12332}{14} = 880.86 \]
\[ s_x = \sqrt{\frac{\sum x^2}{n - 1}} = \sqrt{\frac{130460}{14}} = 96.53 \]
\[ s_y = \sqrt{\frac{\sum y^2}{n - 1}} = \sqrt{\frac{1429.72}{14}} = 10.11 \]

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>550</td>
<td>65</td>
<td>4010</td>
</tr>
<tr>
<td>520</td>
<td>55</td>
<td>2860</td>
</tr>
<tr>
<td>715</td>
<td>65</td>
<td>4645</td>
</tr>
<tr>
<td>405</td>
<td>42</td>
<td>17010</td>
</tr>
<tr>
<td>690</td>
<td>64</td>
<td>43220</td>
</tr>
<tr>
<td>490</td>
<td>45</td>
<td>22750</td>
</tr>
<tr>
<td>565</td>
<td>56</td>
<td>31640</td>
</tr>
<tr>
<td>580</td>
<td>59</td>
<td>34220</td>
</tr>
<tr>
<td>615</td>
<td>58</td>
<td>34480</td>
</tr>
<tr>
<td>455</td>
<td>42</td>
<td>19270</td>
</tr>
<tr>
<td>440</td>
<td>38</td>
<td>16720</td>
</tr>
<tr>
<td>515</td>
<td>50</td>
<td>25750</td>
</tr>
<tr>
<td>590</td>
<td>37</td>
<td>14060</td>
</tr>
<tr>
<td>510</td>
<td>42</td>
<td>21420</td>
</tr>
<tr>
<td>565</td>
<td>53</td>
<td>29410</td>
</tr>
</tbody>
</table>

\[ \Sigma X = 5810 \quad \Sigma Y = 772 \quad \Sigma XY = 424580 \]

\[ \Sigma X^2 = 4407800 \quad \Sigma Y^2 = 41162 \]

CONCLUSIONS

correlation
scattergrams
Pearson r
formulas

EXEMPLARY

covariance formula

\[ r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{880.86}{96.53 \times 10.11} = 0.903 \]

COCLUSION

not just any two variables

EXAMPLE

covariance formula

\[ r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{880.86}{96.53 \times 10.11} = 0.903 \]

where,

\[ s_{xy} = \frac{\sum xy}{n - 1} = \frac{12332}{14} = 880.86 \]
\[ s_x = \sqrt{\frac{\sum x^2}{n - 1}} = \sqrt{\frac{130460}{14}} = 96.53 \]
\[ s_y = \sqrt{\frac{\sum y^2}{n - 1}} = \sqrt{\frac{1429.72}{14}} = 10.11 \]

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>550</td>
<td>65</td>
<td>4010</td>
</tr>
<tr>
<td>520</td>
<td>55</td>
<td>2860</td>
</tr>
<tr>
<td>715</td>
<td>65</td>
<td>4645</td>
</tr>
<tr>
<td>405</td>
<td>42</td>
<td>17010</td>
</tr>
<tr>
<td>690</td>
<td>64</td>
<td>43220</td>
</tr>
<tr>
<td>490</td>
<td>45</td>
<td>22750</td>
</tr>
<tr>
<td>565</td>
<td>56</td>
<td>31640</td>
</tr>
<tr>
<td>580</td>
<td>59</td>
<td>34220</td>
</tr>
<tr>
<td>615</td>
<td>58</td>
<td>34480</td>
</tr>
<tr>
<td>455</td>
<td>42</td>
<td>19270</td>
</tr>
<tr>
<td>440</td>
<td>38</td>
<td>16720</td>
</tr>
<tr>
<td>515</td>
<td>50</td>
<td>25750</td>
</tr>
<tr>
<td>590</td>
<td>37</td>
<td>14060</td>
</tr>
<tr>
<td>510</td>
<td>42</td>
<td>21420</td>
</tr>
<tr>
<td>565</td>
<td>53</td>
<td>29410</td>
</tr>
</tbody>
</table>

\[ \Sigma X = 5810 \quad \Sigma Y = 772 \quad \Sigma XY = 424580 \]

\[ \Sigma X^2 = 4407800 \quad \Sigma Y^2 = 41162 \]

NEXT TIME

factors affecting r
interpreting r
and

Is there a link between IQ and problem solving ability?

CORRELATION

r measures correlation between two variables
not just any two variables

1. The two variables must be paired observations.
2. Variables must be quantitative (interval or ratio scale).