Introduction to Statistics in Psychology
PSY 201
Professor Greg Francis
Lecture 10
correlation

Did I damage my daughter’s eyes?

**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

---

**SCATTERGRAMS**

plot value of one variable against the value of the other variable

---

**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:

CORRELATION COEFFICIENT

quantitative measure of correlation bounded between $-1.0 \leq r \leq +1.0$

- 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

\[ r = \frac{\text{degree to which } X \text{ and } Y \text{ vary together}}{\text{degree to which } X \text{ and } Y \text{ vary separately}} \]

several formulas

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

all give the same result!

PERFECT CORRELATIONS

- perfect positive correlation
- perfect negative correlation

NO CORRELATION

no correlation
balance of larger and smaller values

$z$ 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

**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!

**DEVIATION 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
\[
r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}
\]

**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!

**RAW SCORE FORMULA**

it is awkward to calculate deviation scores
raw score formula
\[
r_{xy} = \frac{n \Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[n \Sigma X^2 - (\Sigma X)^2][(n \Sigma Y^2 - (\Sigma Y)^2)]}}
\]
**Covariance Formula**

\[ \text{covariance} = s_{xy} = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{n - 1} \]

average cross-product of deviation scores

(same to variance)

Pearson r turns out to be:

\[ r = \frac{s_{xy}}{s_X s_Y} \]

where \( s_X \) and \( s_Y \) are the standard deviations of their respective distributions

**Example**

development score formula

\[ r_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{\sum x^2}{n}} \sqrt{\sum y^2 - \frac{\sum y^2}{n}}} = 0.903 \]

<table>
<thead>
<tr>
<th>X</th>
<th>595</th>
<th>520</th>
<th>715</th>
<th>405</th>
<th>680</th>
<th>490</th>
<th>455</th>
<th>589</th>
<th>615</th>
<th>445</th>
<th>545</th>
<th>515</th>
<th>380</th>
<th>540</th>
<th>475</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>68</td>
<td>55</td>
<td>63</td>
<td>31</td>
<td>55</td>
<td>56</td>
<td>56</td>
<td>59</td>
<td>56</td>
<td>56</td>
<td>44</td>
<td>50</td>
<td>37</td>
<td>50</td>
<td>42</td>
</tr>
<tr>
<td>z</td>
<td>0.29</td>
<td>0.22</td>
<td>0.97</td>
<td>2.97</td>
<td>0.45</td>
<td>0.38</td>
<td>1.51</td>
<td>1.94</td>
<td>0.97</td>
<td>1.03</td>
<td>3.64</td>
<td>3.14</td>
<td>1.14</td>
<td>1.03</td>
<td>2.29</td>
</tr>
</tbody>
</table>

\( \sum z_x = 24 \) and \( \sum z_y = 14 \)

\( s_{xy} = 142.92 \)

**Example**

raw score formula

\[ r_{xy} = \frac{\sum xy - \left(\frac{\sum x \sum y}{n}\right)}{\sqrt{\left(\sum x^2 - \frac{\sum x^2}{n}\right) \left(\sum y^2 - \frac{\sum y^2}{n}\right)}} = 0.903 \]

<table>
<thead>
<tr>
<th>X</th>
<th>595</th>
<th>520</th>
<th>715</th>
<th>405</th>
<th>680</th>
<th>490</th>
<th>455</th>
<th>589</th>
<th>615</th>
<th>445</th>
<th>545</th>
<th>515</th>
<th>380</th>
<th>540</th>
<th>475</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>68</td>
<td>55</td>
<td>63</td>
<td>31</td>
<td>55</td>
<td>56</td>
<td>56</td>
<td>59</td>
<td>56</td>
<td>56</td>
<td>44</td>
<td>50</td>
<td>37</td>
<td>50</td>
<td>42</td>
</tr>
<tr>
<td>z</td>
<td>0.29</td>
<td>0.22</td>
<td>0.97</td>
<td>2.97</td>
<td>0.45</td>
<td>0.38</td>
<td>1.51</td>
<td>1.94</td>
<td>0.97</td>
<td>1.03</td>
<td>3.64</td>
<td>3.14</td>
<td>1.14</td>
<td>1.03</td>
<td>2.29</td>
</tr>
</tbody>
</table>

\( \sum z_x = 24 \) and \( \sum z_y = 14 \)

\( s_{xy} = 142.92 \)

\( \text{Covariance} = \sum xy - \left(\frac{\sum x \sum y}{n}\right) \)

**Example**

standard score formula

\[ r_{xy} = \frac{\sum z_x z_y}{n - 1} = \frac{12.67}{14} = 0.905 \]

<table>
<thead>
<tr>
<th>X</th>
<th>595</th>
<th>520</th>
<th>715</th>
<th>405</th>
<th>680</th>
<th>490</th>
<th>455</th>
<th>589</th>
<th>615</th>
<th>445</th>
<th>545</th>
<th>515</th>
<th>380</th>
<th>540</th>
<th>475</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>68</td>
<td>55</td>
<td>63</td>
<td>31</td>
<td>55</td>
<td>56</td>
<td>56</td>
<td>59</td>
<td>56</td>
<td>56</td>
<td>44</td>
<td>50</td>
<td>37</td>
<td>50</td>
<td>42</td>
</tr>
<tr>
<td>z</td>
<td>0.29</td>
<td>0.22</td>
<td>0.97</td>
<td>2.97</td>
<td>0.45</td>
<td>0.38</td>
<td>1.51</td>
<td>1.94</td>
<td>0.97</td>
<td>1.03</td>
<td>3.64</td>
<td>3.14</td>
<td>1.14</td>
<td>1.03</td>
<td>2.29</td>
</tr>
</tbody>
</table>

\( \sum z_x = 24 \) and \( \sum z_y = 14 \)

\( s_{xy} = 142.92 \)
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).

CONCLUSIONS

correlation
scattergrams
Pearson $r$
formulas

NEXT TIME

factors affecting $r$
interpreting $r$
and

Is there a link between IQ and problem solving ability?