Correlation and regression

Rudolf Cardinal & Mike Aitken
11 / 12 November 2004
Department of Experimental Psychology
University of Cambridge

Everything (inc. slides) also at
pobox.com/~rudolf/psychology
Have you read the *Background Knowledge* (§1)?

Did you remember to bring:

- the stats booklet
- your calculator
- your data from the last practical *(mental rotation)*?

If not…

oomp!
Plan of this session:
• we’ll cover the ideas and techniques of correlation and regression. (The handout covers everything you need to know and more: Section 2 for today.)

• you can analyse your own data (± have a go at the examples)

• you can ask questions (about this practical, the background material, or anything statistical) and we’ll try to help.

• Afterwards, practise.
Remember:

Wavy-line stuff in the handout is for reference only.
   You might be interested.
   You might refer to it in the future.
   You do not need to understand or learn it.
You should already know (from NST 1A or booklet §1)...

- measures of **central tendency** (e.g. mean, median, mode)
- measures of **dispersion** (e.g. variance, standard deviation)
- histograms and distributions
- the logic of null hypothesis testing
You should already know (from NST 1A or booklet §1)…

mean
\[ \bar{x} = \frac{\sum x}{n} \]

sample variance
\[ s^2_x = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{\sum x^2 - (\sum x)^2}{n(n-1)} \]

sample standard deviation (SD)
\[ s_x = \sqrt{s^2_x} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - (\sum x)^2}{n(n-1)}} \]
The relationship between two variables
Scatter plots show the relationship between two variables.

A. Positive correlation

B. Negative correlation

C. Zero correlation (and no relationship)

D. Zero correlation (nonlinear relationship)
Correlation

There are no amusing or attractive pictures to do with correlation anywhere in the world.
The covariance measures how much two variables vary together. Good name, eh?

Sample covariance:

\[
\text{COV}_{XY} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}
\]

There’s another formula (below) that’s easier to use if you have to calculate the covariance by hand. But you shouldn’t need to if you can work your calculator, because... (next slide...)

\[
\text{cov}_{XY} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n - 1}
\]
The covariance is big and positive if there is a strong positive correlation, big and negative if there is a strong negative correlation, and zero if there’s no correlation. But how big is ‘big’? That depends on the standard deviations (SDs) of X and Y, which isn’t very helpful. So instead we calculate $r$: 

$$r_{XY} = \frac{\text{COV}_{XY}}{S_X S_Y}$$

$r$ varies from $-1$ to $+1$. Your calculator calculates $r$. $r$ does not depend on which way round $X$ and $Y$ are.
**WORKED EXAMPLE: mental rotation**

**Work through this one now with your calculator.**
We’ll pause for a moment to do that...

<table>
<thead>
<tr>
<th>Angle</th>
<th>0</th>
<th>60</th>
<th>120</th>
<th>180</th>
<th>240</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group mean RT (ms, to 0 d.p.)</td>
<td>830</td>
<td>908</td>
<td>1079</td>
<td>1387</td>
<td>1070</td>
<td>935</td>
</tr>
</tbody>
</table>

*(simplest task)*

For correlation ($r$):

Enter linear regression (LR) mode

Clear the stats memory

Enter values of $x$, $y$ pairs (e.g. $x = 53$, $y = 17$)

Read out desired coefficients *(see keypad and inside lid)*

**Casio fx115s**

- **MODE**
- **3**
- **SHIFT**
- **Scl**
- **C**
- **[---]**
- **5**
- **3**
- **1**
- **7**
- **M+**
- **DATA DEL**
- **r**
- **9**
- \(x_{\bar{D}} y_{\bar{D}}\)
- **RCL**
- **3**

**Other Casio models**

- **MODE**
- **→REG→Lin**
- **SHIFT**
- **Scl**
- **AC**
- **=**
- **5**
- **3**
- **1**
- **7**
- **M+**
- **etc.**
- **SHIFT**
- **r**
- **(**
- **=**
- **C**
- **RCL**
- **hyp**
**Oops.**
Not very linear.

*Shortest* angle (so ‘240°’ becomes 120°; ‘300°’ becomes 60°).
Much more linear.
Never give up! Never surrender! Never forget...

‘Zero correlation’ doesn’t imply ‘no relationship’.

So always draw a scatter plot.

Correlation does not imply causation.
Correlation, causation... A real-world example

- Male college students (18–36 years old) engaged in 5 min conversation with a stooge, who was either **male** (23 or 32 y.o.) or **female** (19–23 y.o.). Didn’t know that this was part of an experiment.

- Saliva samples taken before and after conversation. Testosterone (T) measured.

- ‘Recent sexual experience’ = current relationship or sex in last 6 months.

- Stooge rated how much they thought the subject was ‘displaying’ to them (e.g. talkative, showed off, tried to impress).

*Roney et al. (2003)*
Adjusted $r$

If we sample only 2 $(x, y)$ points, we’ll get a perfect correlation in the sample, $r = 1$ (or $-1$). But that doesn’t mean that the correlation in the underlying population, $\rho$, is perfect! A better estimator of $\rho$ is $r_{adj}$:

$$r_{adj} = \sqrt{1 - \frac{(1 - r^2)(n-1)}{n-2}}$$
Adjusted $r$: mental rotation example

\[ r = 0.954 \]

\[ n = 6 \]

\[ r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = 0.942 \]
Beware of sampling a restricted range
Beware outliers!
Heterogeneous subgroups: the oncologist and the magic forest

**J**
- **mass (kg)**
- **height (cm)**
- *wild boar*
- *runner-bean plants*

**K**
- **risk of cancer**
- **exposure to chemical**
- *vulnerable subgroup*
- *resistant subgroup*
‘Is my correlation significant?’ Our first $t$ test.

**Research hypothesis:** the correlation in the underlying population from which the sample is drawn is non-zero ($\rho \neq 0$). **Null hypothesis:** the correlation in the population is zero ($\rho = 0$).

Calculate $t$:

$$t_{n-2} = \frac{r \sqrt{n-2}}{\sqrt{1 - r^2}}$$

A big $t$ (positive or negative) means that your data would be unlikely to be observed if the null hypothesis were true. Look up the **critical level** of $t$ for “$n–2$ degrees of freedom” in the *Tables and Formulae*. Values of $t$ near zero are not ‘significant’. If your $|t|$ is more extreme than the critical value, it is significant. If your $t$ is significant, you **reject** the null hypothesis. Otherwise, you don’t. We’ll explain $t$ tests properly in the next practical.
t test: mental rotation example

\[ r = 0.954 \]
\[ n = 6 \]
\[ t_{n-2} = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}} = 6.36 \]

Critical \( t_4 = 2.776 \) for \( \alpha = 0.05 \) two tailed
Assumptions you make when you test hypotheses about $\rho$

Basically, the data shouldn’t look too weird. We must assume

- that the variance of $Y$ is roughly the same for all values of $X$. (This is termed homogeneity of variance or homoscedasticity. Its opposite is heterogeneity of variance or heteroscedasticity.)

- that $X$ and $Y$ are both normally distributed

- that for all values of $X$, the corresponding values of $Y$ are normally distributed, and vice versa
Heteroscedasticity is a Bad Thing (and Hard to Spell)

This would violate an assumption of testing hypotheses about $\rho$. So if you run the $t$ test on $r$, the answer may be meaningless.
$r_s$: Spearman’s correlation coefficient for ranked data

- Rank the $X$ values.
- Rank the $Y$ values.
- Correlate the $X$ ranks with the $Y$ ranks. (You do this in the normal way for calculating $r$, but you call the result $r_s$.)

- To ask whether the correlation is ‘significant’, use the table of critical values of Spearman’s $r_s$ in the *Tables and Formulae* booklet.
How to rank data

Suppose we have ten measurements (e.g. test scores) and want to rank them. First, place them in ascending numerical order:

<table>
<thead>
<tr>
<th>X</th>
<th>5</th>
<th>8</th>
<th>9</th>
<th>12</th>
<th>12</th>
<th>15</th>
<th>16</th>
<th>16</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
</table>

Then start assigning them ranks. When you come to a tie, give each value the mean of the ranks they’re tied for — for example, the 12s are tied for ranks 4 and 5, so they get the rank 4.5; the 16s are tied for ranks 7, 8, and 9, so they get the rank 8:

<table>
<thead>
<tr>
<th>X:</th>
<th>5</th>
<th>8</th>
<th>9</th>
<th>12</th>
<th>12</th>
<th>15</th>
<th>16</th>
<th>16</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>rank:</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4.5</td>
<td>4.5</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>
Regression
Linear regression: predicting things from other things

\[ \hat{Y} = a + bX \]

But which line? Which values of \( a \) and \( b \)?
‘Least squares’ regression: finding \( a \) and \( b \)

Predicted value of \( Y \):

\[
\hat{y} = a + bx
\]

Error in prediction (residual):

\[
y - \hat{y}
\]

Squared error:

\[
(y - \hat{y})^2
\]

Sum of squared errors:

\[
\sum (y - \hat{y})^2
\]

We pick values of \( a \) and \( b \) in such a way that minimizes the sum of the squared errors.
We’ve found our best line.

\[ \hat{Y} = a + bX \]

\[ a = \overline{y} - b \overline{x} \]
\[ b = \frac{\text{COV}_{XY}}{s^2_X} = r \frac{s_Y}{s_X} \]

Our line will always pass through \((0, a)\) and \((x, \overline{y})\).
Beware extrapolating beyond the original data
Predicting Y from X is **not** the same as predicting X from Y!

**A**  
\[ \hat{Y} = a + bX \]

*Predicting Y from X.*  
Residuals: the difference between each point and the predicted value.

**B**  
But the best line for predicting Y from X may be very bad at predicting X from Y...

**C**  
This is a much better line for predicting X from Y; the residuals are much smaller on average.

**D**  
\[ \hat{X} = a_2 + b_2Y \]

\[ \hat{Y} = a + bX \]

*Both are lines of best fit that minimize the sum of the squares of the residuals.*
$r^2$ means something important — the proportion of the variability in $Y$ predictable from the variability in $X$
Doing it for real
For **descriptive statistics** (mean, SD, etc.):  

Enter descriptive statistics (SD) mode

Clear the stats memory

Enter values of $x$  
(e.g. 53)

Read out descriptive statistics:  
(mean  
(see keypad and inside lid)  
(sample SD ($n-1$ formula))  
$n$

For correlation and **linear regression** ($r, a, b$):  

Enter linear regression (LR) mode

Clear the stats memory

<table>
<thead>
<tr>
<th>shortest angle ($^\circ$)</th>
<th>RT (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>830</td>
</tr>
<tr>
<td>60</td>
<td>908</td>
</tr>
<tr>
<td>120</td>
<td>1079</td>
</tr>
<tr>
<td>180</td>
<td>1387</td>
</tr>
<tr>
<td>120</td>
<td>1070</td>
</tr>
<tr>
<td>60</td>
<td>935</td>
</tr>
</tbody>
</table>

Enter values of $x, y$ pairs  
(e.g. $x = 53$, $y = 17$)

Read out desired coefficients  
(see keypad and inside lid)
Mental rotation: regression line (group means, simplest task)

\[ y = 765 + 2.995x \]

\[ r^2 = 0.910 \]

\[ r = 0.954 \]
Thu 11 November 2004: Armistice Day