Lecture #16:

Correlation and Regression

\[ y = \text{tail length (cm)} \]
\[ x = \text{wing length (cm)} \]

**Sample**

\[
\begin{array}{c|c}
7.4 & 10.4 \\
7.8 & 10.8 \\
\vdots & \vdots \\
8.3 & 10.4 \\
\end{array}
\]
\[ n = 12 \]

\[
\bar{y} = 27.4 \quad \bar{x} = 10.7 \text{ cm}
\]

SD \( s_y = 0.35 \quad SD s_x = 0.43 \text{ cm} \)

**Positive association (linear)**

**Scatterplot (scatter diagram)**

Typical w/ 2 normal curves

Elliptical shape (bivariate normal)

Most scatterplots we see in class

**Negative association (linear)**

**No association (linear)**

- Karl Pearson 
- Edward Francis Galton (1875)

- **L-216**
Correlation between X and Y:

\[ r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \]

-225

Correlation coefficient is always in "pure number" without units.

\[ s_{xy} = \sqrt{\sum (y_i - \bar{y})(x_i - \bar{x})^2} / n \]

- Data

- Error

- Data

- Error

- Positive slope

- Negative slope

- Hormone level

- Age

- Y = 0

- r = 1

- y can be fooled...

- Or is always in "pure number" without units.

- Positive slope

- Negative slope

- Area where more points are found

- Area where less points are found

- Above y = 0, value of y

- Below y = 0, value of y

- Point of averages

- Above y = 0, value of y

- Below y = 0, value of y

- Above y = 0, value of y

- Below y = 0, value of y
- Unusual to x but not y
- Unusual to y but not x

Univariate outlier

Bivariate outlier

R-73 - Page of scatterplots
- Training your eye to read correlation values

\[
\begin{array}{c|c}
 y & x \\
\hline
 3 & 3 \\
 9 & 9 \\
-4 & -4 \\
\end{array}
\]

\[r = 1\]

- Here \( r = 0.87 \) is a strong but not perfect linear attraction between width and tail length

4. If you add a constant to all x or y values, y is unchanged (just change location), bc so is unchanged

5. If you multiply a constant to all x or y values, \( r \) is unchanged, or if a neg constant, \( r \) changes sign

Q. Is an \( r \) of \( +0.87 \) large in practical terms? (is it practical?)
A: A sparrow w/ smallest $x$ (≈10 cm) has $y ≈ 7.25$ cm. A sparrow with largest $x ≈ 11.25$ and $y ≈ 8$ cm - 7.5 and 8 are dramatically different in practical (sharply different from 0)

Q: IS $r = +0.87$ statsig?
A: Build inferential model → L-229

Inferential summary

<table>
<thead>
<tr>
<th>Unknown pop q of interest</th>
<th>$p$ = pop corr between wing and tail length in this species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate of $p$</td>
<td>$r = +0.87$</td>
</tr>
<tr>
<td>Give or fake for $95%$</td>
<td>$s_{p} = \sqrt{(1 - r^2) / n}$ = 0.081</td>
</tr>
<tr>
<td>CI for $p$</td>
<td>approx (0.71, 1.0) exact (0.59, 0.96)</td>
</tr>
</tbody>
</table>

L-231 → L-244

Extra credit material

L: Takehome final - formula on R-25

$+0.87 \pm (2)(0.081)$

$0.71 \leq 0.87 \leq 0.98$

A: Diff is statsig!