Last time we learned about scatter plots and how they provide a visual assessment of the relationship between 2 continuous variables.

Now, how do we quantify the strength of the association between 2 continuous variables?

There are many measures of strength (like there are different measures of spread of a histogram), and each one captures a different facet of strength. One popular measure is Pearson's correlation coeff., denoted $r$ (for sample) and $\rho$ (for distribution), i.e. population

$r$ gives a point estimate of $\rho$.

$\bar{x}$ gives a point estimate of $\mu_x$.

\[ \uparrow \quad \uparrow \]

Sample mean

Population mean
How do we compute it?

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

\[ r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \]

later!

= Average of "areas" →

Important: The specific measure of strength that r measures is the "skinniness" of the scatterplot.

generally \{ fat scattering → r ≈ 0 \} skinny, → r ≈ 1, But
Important: $r$ is a summary measure of a scatterplot. As such, some info is lost when you look only at $r$. Look at the scatterplot (too)!

- $r = +1$
- $r = -1$
- $r = 0$
- $r \sim 0.6 - 0.7$
- $r \sim -0.6, 0.7$
- $r \sim 0$
1. How does switching \( x \) and \( y \) affect \( r \)?

\[ \text{Because } r_{xy} = r_{yx} \]

(See Lab)

2. How does scaling (i.e. multiplying all \( x \) or \( y \) values by some number) affect \( r \)?

It does not!

\( r \) is invariant under scaling.

Because \( z_i = \frac{x_i - \bar{x}}{s_x} \)

\[ \begin{align*}
\bar{z}_i &= \frac{\sum (c x_i - c \bar{x})}{n-1} \\
&= \frac{c \sum (x_i - \bar{x})}{n-1} \\
&= c \frac{\sum (x_i - \bar{x})}{n-1} \\
&= c \frac{S_x}{S_x} \\
&= c S_x
\end{align*} \]

How about shifting \( x \) or \( y \)?

Show!

\[ x_i \rightarrow x_i + c \quad \text{or} \quad y_i \rightarrow y_i + c \]

Generally, \( r \) has the following properties:

- \(-1 \leq r \leq 1\) \quad \text{"simplicity"}
- \( r_{xy} = r_{yx} \)
- Measure of linear assoc. \( \text{spread about line} \)
- Unaffected by scaling, shifting, ...
- Misleading! (Next time!)
Very important: Relation $\not\Rightarrow$ Causation.

Even if there is a strong (linear) relationship between 2 variables, that does not mean that one causes the other.

Shoe size and reading ability are correlated.

But even an acausal relationship can be used for predicting one from the other.

You can predict reading ability from shoe size.
We have learned that \( r \) (Pearson's correlation coefficient) is a measure of the strength of the linear relationship between 2 continuous variables, with "strength" measured by "skinniness". But \( r \) can be misleading.

When you see \( r = \text{large} \) (e.g., 0.9) or \( r = \text{small} \) (0.1), you should wonder if \( r \) is lying to you.

- **There are situations which make \( r \) "artificially" small:**

  1) When there is a nonlinear rel.,
  2) When there are outliers,
  3) When there are clusters.

Also keep in mind that \( r \neq \varphi \) even if \( r = 0.9 \), \( \varphi \) may still be 0. And vice versa.

- **There are situations which make \( r \) "artificially" large:**

Also, " ecological correl" in lab.

**Moral:** \( r \) is misleading if the scatterplot has clusters, outliers, ... So, regardless of the \( r \) value you get in your problem, look at the scatterplot too.
I gave you a formula that defines $r$. (The book gives you 2 others on p. 108.)

a) Start from the formula I gave you, and show that it is equal to

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$  \(\text{I}\)

b) Start from (I), and show that it is equal to

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

with $S_{xx}$, $S_{xy}$, $S_{yy}$ defined on p. 108.

Suppose $n$ cases of data on $x$ and $y$ fall exactly on the line $y = mx + b$. Compute the value of $r$.

Hint: In any of the formulas for $r$, eliminate all $y$ in favor of $x$.

The $z$'s appearing in the formula for $r$ have two nice properties: Their sample mean is zero, and their sample variance is 1. Prove these!

I.e. show $\frac{1}{n} \sum z_i = 0$, $\frac{1}{n-1} \sum (z_i - \bar{z})^2 = 1$