Thus far, our focus has been on 1 column of data, and 1 variable. I.e. univariate analysis.

With 2 (or more) variables, we can do all of the above, but we can also ask about the relationship between them.

For continuous data: scatterplot.

The scatterplot is "the best" way of visualizing the association (relationship) between 2 continuous variables.

Not unusual. In fact, they are common, (and even necessary)
Scatter plot

Museum?

Random association
no association/relation

\( \text{weight} \)
\( \text{height} \)

linear, constant variance
\( y \) generally increases with \( x \)
but \( y \)'s variance does not.

linear, not-constant variance.
As \( x \) increases, \( y \) generally increases, but the var. of \( y \) generally increases, too.

nonlinear, but monotonic.

nonlinear and non-monotonic
\( y \) generally decreases with increasing \( x \),
but only up to some point, and
then generally increases with \( x \).

When you see 2 columns of data on continuous variables,
don't think (!), just do the scatterplot. Then think.

Of course, don't forget to look at the hist of each
How can we quantify the association seen in the scatterplot?

There are many measures of association. The same way there are many measures of “center” or “spread” of hists. They measure different facets of the “strength” of association. One popular measure is Pearson’s correlation coeff. denoted \( r \) (for sample) and \( \rho \) (for population). Like \( \bar{x}, \bar{y}, \mu, \sigma \).

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

\(-1 \leq r < +1\)

**Important:** \( r \) measures “skinniness” of scatterplot.

Fat scatterplot \( \Rightarrow r \approx 0\)

Skinny \( \Rightarrow r \approx \pm 1\).

**But, there are exceptions** (below)

1) \( r = \text{Average of } \overline{\epsilon} \).

Only FYI, do not use on hw/tests.

2) \( r \approx \frac{\bar{x} \cdot \bar{y}}{|x| y |} \sim \text{Cor}(\theta)\)

(Recall: \( s^2 \sim \frac{S^2}{\bar{X} \cdot \bar{Y}}\) n-vector

\[
\bar{x} = (x_1, x_2, x_3, \ldots x_n)
\]

\[
\bar{y} = (y_1, y_2, y_3, \ldots y_n)
\]
Examples (r: museum)

\[ \begin{align*}
\text{45}\degree & \quad r = +1 \\
\text{89}\degree & \quad r = +1 \\
\text{r} & \sim 0.7, 0.8 \\
\text{r} & \sim 0.6 \\
\text{r} & \sim 0
\end{align*} \]

\[ \Rightarrow r \text{ is a measure of linear association} \]

Finally, there are some (unrealistic) situations with \( r = 0 \):

\[ \begin{align*}
\text{Y} & \uparrow \text{horiz.} \\
\text{Y} & \uparrow \text{vertical} \\
\text{X} & \downarrow
\end{align*} \]

proving these requires taking some careful limits. But they are not important.

Important: \( r \) is a summary measure (of skinniness) of a scatterplot. As such, information in the scatterplot (a 2D Thing) is lost when it's reduced to a single number (a 1D Thing). So, always check the scatterplot!
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 $z_i = \frac{x_i - \bar{x}}{S_x}$, $s_y = \sqrt{\frac{1}{n-1} \sum (y_i - \bar{y})^2}$
  - $\bar{y} = \frac{1}{n} \sum y_i$

How does switching $x$ and $y$ affect $r$?

- $r = \frac{1}{n} \sum z_x z_y \rightarrow \frac{1}{n} \sum z_y z_x = r$
  - So, $r$ does not change if $x$ and $y$ are switched.

Many more properties (move, below, linear), but so far:

1) $-1 \leq r \leq +1$
2) is blind to non-linear relations.
3) $r_y = r_{yx}$
4) it's unaffected by scaling (shifting)
BUT, every summary measure (including \( r \)) can mislead.

When you see \( r \) large (e.g. 0.9) or \( r \) 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 see “ecological correl” in Lab.

Moral: \( r \) (like any other summary measure) can be misleading if the data have clusters, outliers, ... So, regardless of the \( r \) value you get in your problem, look at the scatter plot, too.
Make a scatterplot of the 2 continuous vars in the data you collected (ByR). Describe the relationship. Sometimes, taking square-root, or log, of each variable can help in getting a more linear relationship. If it can't be done, see me!

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 $y$'s in favor of $x$'s.

I gave you a formula that defines $r$. The book gives two others on p. 110.

a) Start from the formula I derived in class, and show that it is equal to

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

b) In the lecture where we "derived" the formula for the sample variance, we introduced $S_{xx} = \sum (x_i - \bar{x})^2$, i.e. the numerator of sample variance. If we similarly define $S_{yy}$ and $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$, then it follows from (1) that

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

(which is the other formula for $r$ on p. 110).

To finish all the comparison between my formula for $r$ and those in the book, show that $S_{xy}$ as defined here is equal to

$$ \sum xi yi - n \bar{x} \bar{y} $$

(or $n(\bar{xy} - \bar{x}\bar{y})$) as written in the book.

Hint: Recall $\sum (y_i - \bar{y}) = 0$ [or $\sum (x_i - \bar{x}) = 0$].