Correlation

Suppose:

- A population of N individuals, each having values of X and Y (e.g., height and weight). Think of N as extremely large.
- A random sample of size n individuals from this population, with values $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.

The **sample correlation** r between X and Y is defined by:

$$r = \frac{c(X, Y)}{s_x s_y}$$

where $c(X, Y)$ is the **sample covariance**:

$$c(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

and s_x, s_y are the **sample standard deviations** of X, Y:

$$s_x = \sqrt{s_x^2}\,, \quad s_y = \sqrt{s_y^2}$$

which are the square roots of the **sample variances**:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\,, \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\,,$$

where (of course) $\overline{X}$ and $\overline{Y}$ are the **sample means**:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

All of the previous "sample" quantities give us estimates of the corresponding "population" quantities which are defined by similar summations over the entire population, changing every occurrence of $n-1$ (or n) to N.

$$
\begin{array}{cc}
\underline{\text{sample}} & \underline{\text{population}} \\
\overline{X} & \mu_x = EX \\
\overline{Y} & \mu_y = EY \\
s_x^2 & \sigma_x^2 = E(X-\mu_x)^2 \\
s_y^2 & \sigma_y^2 = E(Y-\mu_y)^2 \\
c(X,Y) & \text{Cov}(X,Y) = E(X-\mu_x)(Y-\mu_y) \\
r = \dfrac{c(X,Y)}{s_x s_y} & \rho = \dfrac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}
\end{array}
$$

In the above, E denotes an "expected value", essentially a "population average", an average of the quantity of interest over all individuals in the population.

The population values listed above are typically unknown, but if you have a large enough "simple" random sample, the *sample* values above will be approximately equal to the *population* values. (Sometimes this is true for other kinds of samples too.)

Interpretation of the Correlation

Range of possible values: $-1 \leq \rho \leq 1\,, \quad -1 \leq r \leq 1$

The population correlation ρ is a measure of the strength of the linear relationship between X and Y in the population.

Similarly, the sample correlation r is a measure of the strength of the linear relationship between X and Y in the sample.

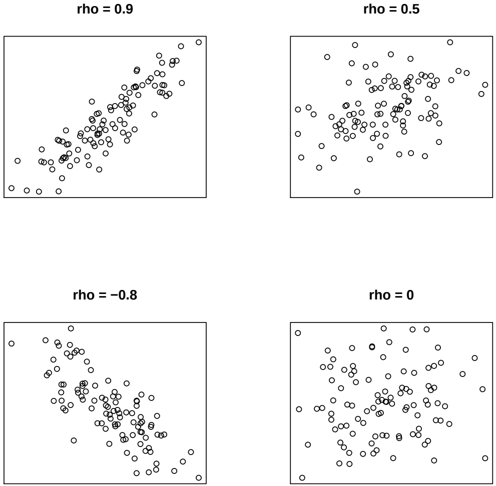
$r = \pm 1$ means there is a perfect linear relationship between X and Y in the sample.

If $r = +1$, the n points $(X_1, Y_1), (X_2, Y_2,), \ldots (X_n, Y_n)$ in the sample lie exactly on a straight line with positive slope.

If $r = -1$, the n points lie exactly on a straight line with negative slope.

(Similar statements hold for the population if $\rho = \pm 1$.)

Scatterplots of samples of size $n = 100$ from populations with different correlations $\rho =$ rho.



Uncorrelated versus Independent

If $\rho = 0$ (equivalently $\text{Cov}(X, Y) = 0$), we say X and Y are **uncorrelated**.

$\rho = 0$ means (roughly) that there is no **linear** relationship between X and Y.

If X and Y are **independent**, this means (roughly) there is **no relationship** between X and Y **of any kind**. The value of X tells us nothing about the value of Y (and vice versa).

If X and Y are **not independent**, we call them **dependent**.

Fact: If X and Y are independent, they are also uncorrelated.

But not the other way around! Uncorrelated X and Y may fail to be independent.

In many cases, uncorrelated X and Y are roughly independent, but not always!

Example: Here are scatterplots of two samples of size $n = 100$ drawn from two different populations (call them L and R for Left and Right). Both populations have $\rho = 0$. (X and Y are uncorrelated.)

X and Y are independent in population L, but **not** in population R. In population R, there is a strong quadratic relation between X and Y.