

Lecture 36 Exploring Bivariate Data Using Correlation

BIO210 Biostatistics

Xi Chen

Fall, 2023

School of Life Sciences

Southern University of Science and Technology



南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

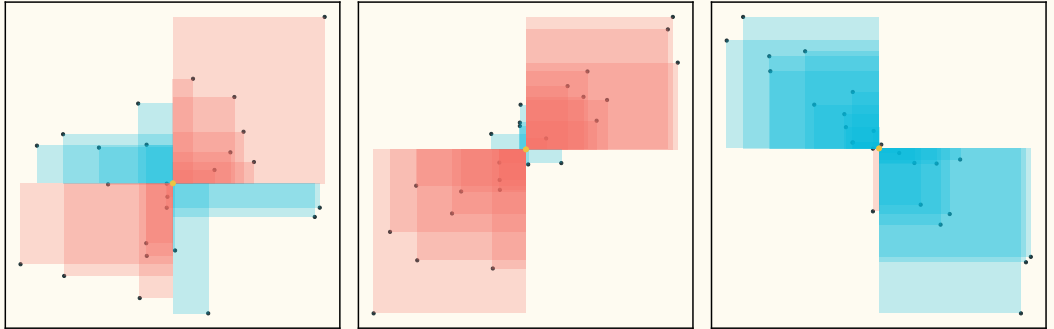
Covariance

$$\begin{aligned}\sigma(\mathbf{X}, \mathbf{Y}) &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}]) \cdot (\mathbf{Y} - \mathbb{E}[\mathbf{Y}])] \\&= \mathbb{E}[\mathbf{XY} - \mathbf{X} \cdot \mathbb{E}[\mathbf{Y}] - \mathbf{Y} \cdot \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{Y}]] \\&= \mathbb{E}[\mathbf{XY}] - \mathbb{E}[\mathbf{X} \cdot \mathbb{E}[\mathbf{Y}]] - \mathbb{E}[\mathbf{Y} \cdot \mathbb{E}[\mathbf{X}]] + \mathbb{E}[\mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{Y}]] \\&= \mathbb{E}[\mathbf{XY}] - \mathbb{E}[\mathbf{Y}] \cdot \mathbb{E}[\mathbf{X}] - \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{Y}] + \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{Y}] \\&= \mathbb{E}[\mathbf{XY}] - \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{Y}]\end{aligned}$$

$$Cov(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

If \mathbf{X} and \mathbf{Y} are independent: $\sigma(\mathbf{X}, \mathbf{Y}) = 0$

Visualisation of The Covariance



From [stats.StackExchange.com](https://stats.stackexchange.com)

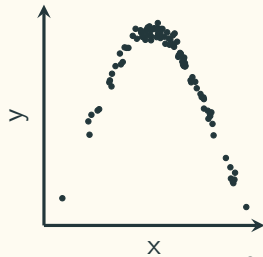
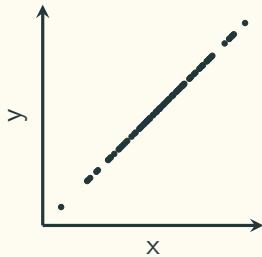
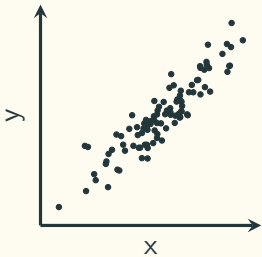
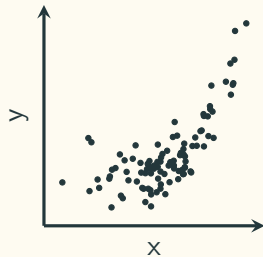
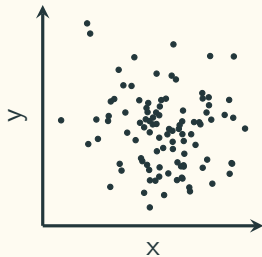
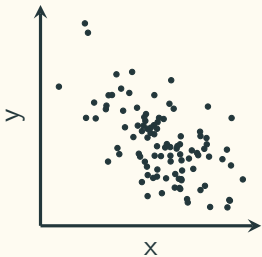
Scatter Plot

The same subject $\left\{ \begin{array}{l} \text{variable 1} \\ \text{variable 2} \end{array} \right.$

Person

var. 1 (weight)	var 2. (height)
--------------------	--------------------

x	x
x	x
x	x
\vdots	\vdots

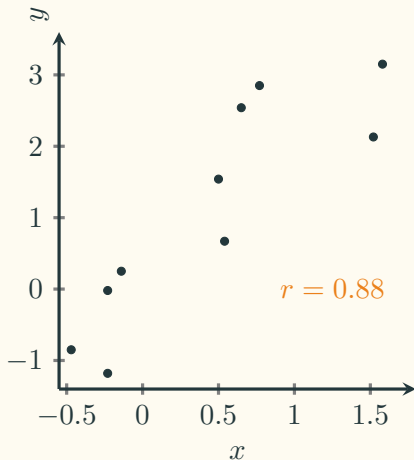


Pearson's Correlation Coefficient (r)

x	y
0.5	1.54
-0.14	0.25
0.65	2.54
1.52	2.13
-0.23	-1.18
-0.23	-0.02
1.58	3.15
0.77	2.85
-0.47	-0.85
0.54	0.67

$$\bar{x} = 0.45, \bar{y} = 1.11$$

$$s_x = 0.72, s_y = 1.55$$



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

$$r = \frac{1}{n-1} \sum_n Z_{x_i} Z_{y_i}$$

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right]}} \\ &= \frac{Cov(x, y)}{\sqrt{Cov(x, x) \cdot Cov(y, y)}} \end{aligned}$$

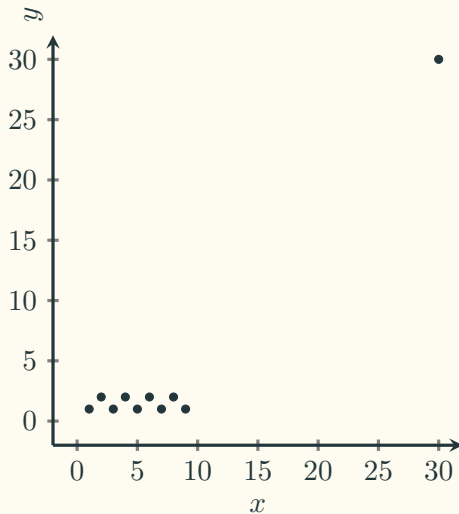
$$-1 \leq r \leq 1$$

Pearson's Correlation Coefficient (r)

x	y
1	1
2	2
3	1
4	2
5	1
6	2
7	1
8	2
9	1
30	30

$$\bar{x} = 7.5, \bar{y} = 4.3$$

$$s_x = 8.32, s_y = 9.04$$

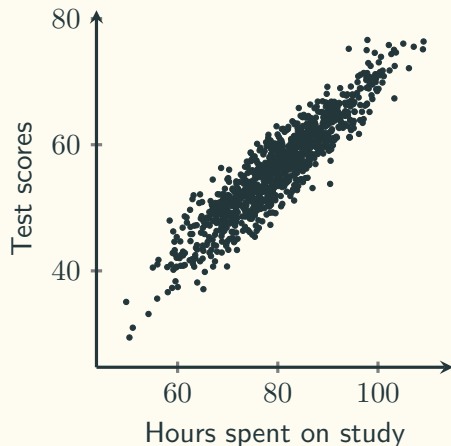


$$r = 0.95$$

**Be careful about
outliers!**

Hypothesis testing of Pearson's r

We suspect that there is a linear relationship between the number of hours spent on study and the test scores. To find out if this is the case, we can draw a random sample and conduct a hypothesis testing.



Population correlation coefficient: ρ

Sample correlation coefficient: r

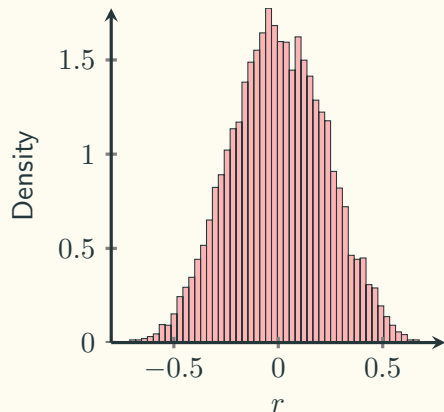
$$\begin{cases} H_0 : \text{no linear relationship} \\ H_1 : \text{some linear relationship} \end{cases} \Leftrightarrow \begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

What is the sampling distribution of r ?

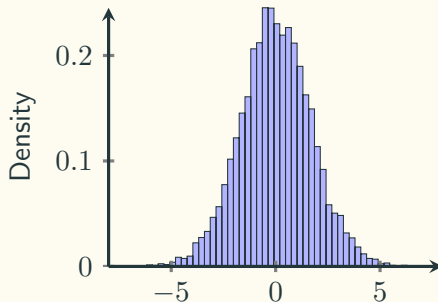
Sampling Distribution of Pearson's r

Under H_0 (no linear relationship) is true:

10,000 simulations under H_0 is true



$$r \sqrt{\frac{n-2}{1-r^2}} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} \sim t_{n-2}$$



Hypothesis testing of Pearson's r

To investigate whether there is a linear relationship between the number of hours spent on study and the test scores, 20 students were randomly selected, and Pearson's r was calculated to be $r = 0.69$.

$$\text{Test statistic: } t = r \sqrt{\frac{n-2}{1-r^2}} = 0.69 \times \sqrt{\frac{20-2}{1-0.69^2}} = 4.04$$

$$\text{Two-tailed } p\text{-value: } \mathbb{P}(|t| \geq 4.04) = 2 \times \mathbb{P}(t \geq 4.04) = 0.000768$$