Lecture 36 Exploring Bivariate Data Using Correlation

BIO210 Biostatistics

Xi Chen

Spring, 2022

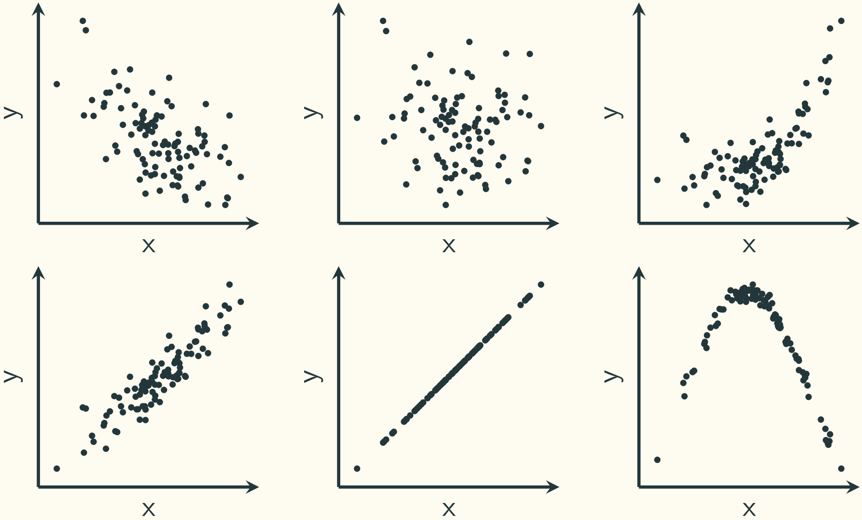
School of Life Sciences
Southern University of Science and Technology

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

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
⋮	⋮

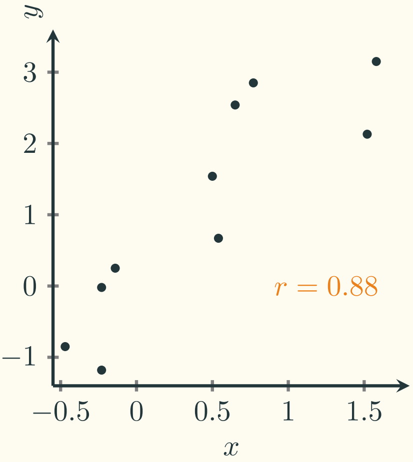


Pearson's Correlation Coefficient (r)

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

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

$s_x = 0.72,\ s_y = 1.55$



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

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

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum\limits_{i=1}^{n} (x_i - \bar{x})^2 \right] \left[\sum\limits_{i=1}^{n} (y_i - \bar{y})^2 \right]}}$$

$$-1 \leqslant r \leqslant 1$$

Covariance

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

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

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

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

Variance of The Sum of Two Random Variables

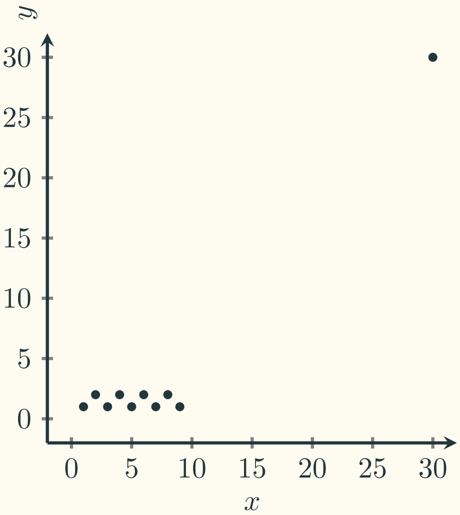
$$
\begin{aligned}
var(\boldsymbol{X}+\boldsymbol{Y}) &= E[(\boldsymbol{X}+\boldsymbol{Y})^2] - (E[\boldsymbol{X}+\boldsymbol{Y}])^2 \\
&= E[\boldsymbol{X}^2 + 2\boldsymbol{X}\boldsymbol{Y} + \boldsymbol{Y}^2] - (E[\boldsymbol{X}] + E[\boldsymbol{Y}])^2 \\
&= E[\boldsymbol{X}^2] + 2\cdot E[\boldsymbol{X}\boldsymbol{Y}] + E[\boldsymbol{Y}^2] - (E[\boldsymbol{X}])^2 - 2\cdot E[\boldsymbol{X}]E[\boldsymbol{Y}] - (E[\boldsymbol{Y}])^2 \\
&= \Big(E[\boldsymbol{X}^2] - (E[\boldsymbol{X}])^2\Big) + \Big(E[\boldsymbol{Y}^2] - (E[\boldsymbol{Y}])^2\Big) + 2\Big(E[\boldsymbol{X}\boldsymbol{Y}] - E[\boldsymbol{X}]E[\boldsymbol{Y}]\Big) \\
&= var(\boldsymbol{X}) + var(\boldsymbol{Y}) + 2\cdot\sigma(\boldsymbol{X},\boldsymbol{Y})
\end{aligned}
$$

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

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

$s_x = 8.32,\ s_y = 9.04$

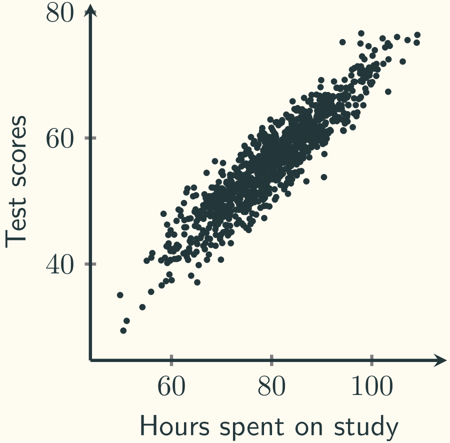


$r = 0.95$

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

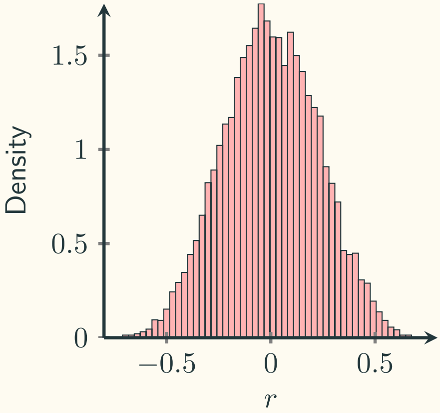
Sample correlation coefficient: r

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

What is the sampling distribution of $\boldsymbol{r}$?

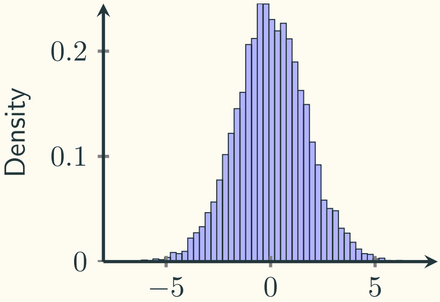
Sampling Distribution of Pearson's r

10,000 simulations under H_0 is true



Under H_0 (no linear relationship) is true:

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



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 $r = 0.69$.

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

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