Statistics with Spa R R ows

Lecture 10

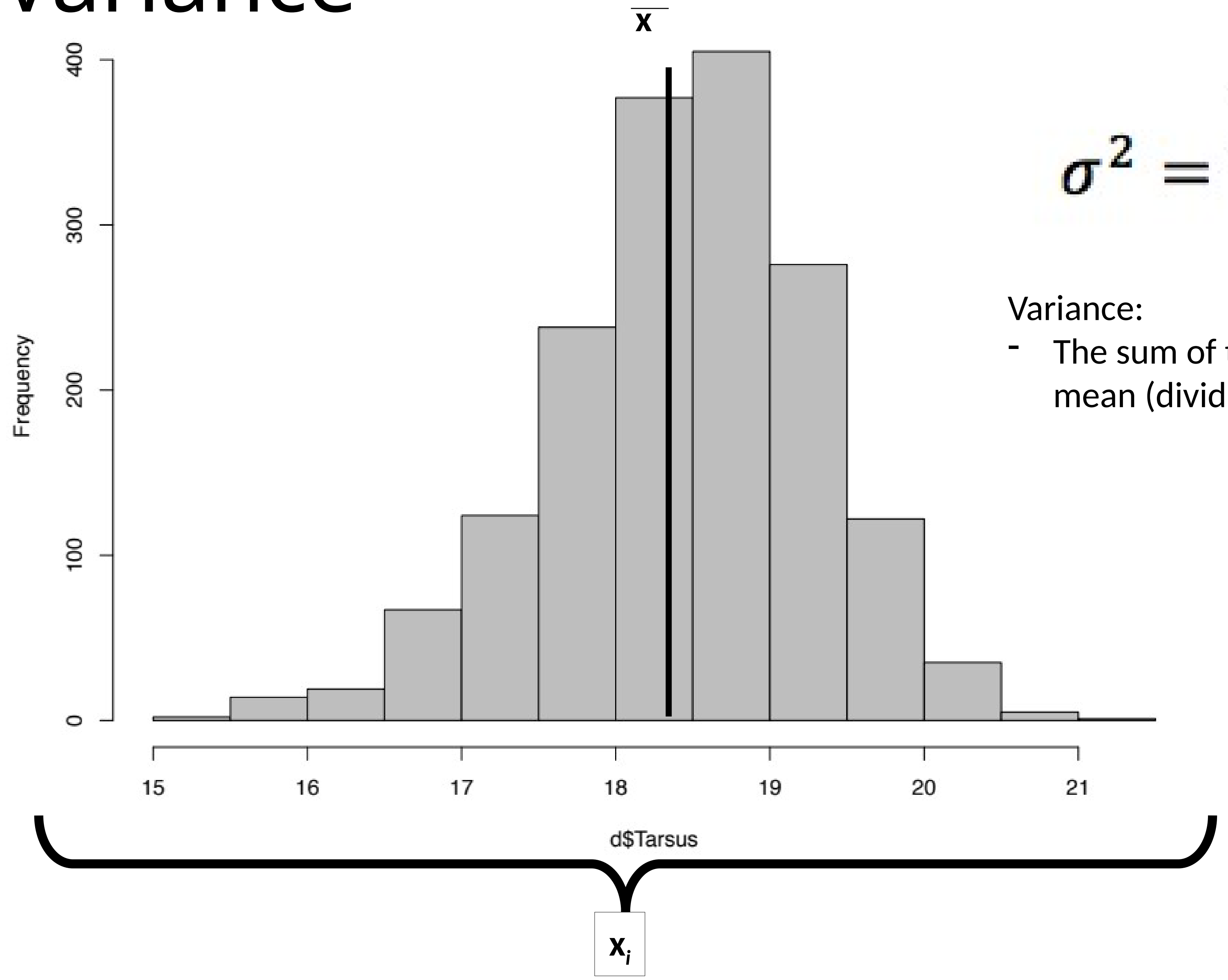
Julia Schroeder

Julia.schroeder@imperial.ac.uk

Aims

- Understanding covariance
- R^2
- Understanding correlations
- Sums of squares

Variance



$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

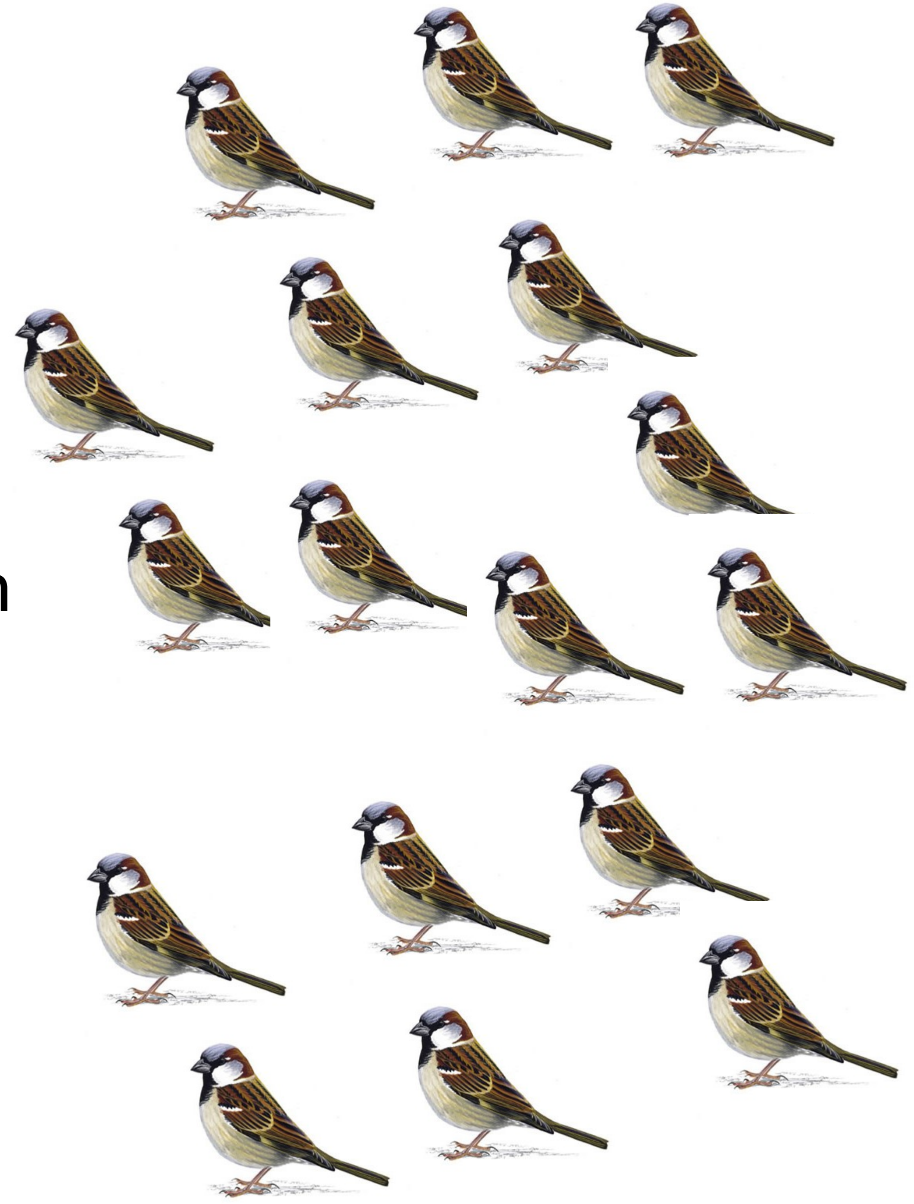
Variance:

- The sum of the squared deviations from the mean (divided by n-1)

Covariance

- How two variables change together

- Population: joint probability distribution

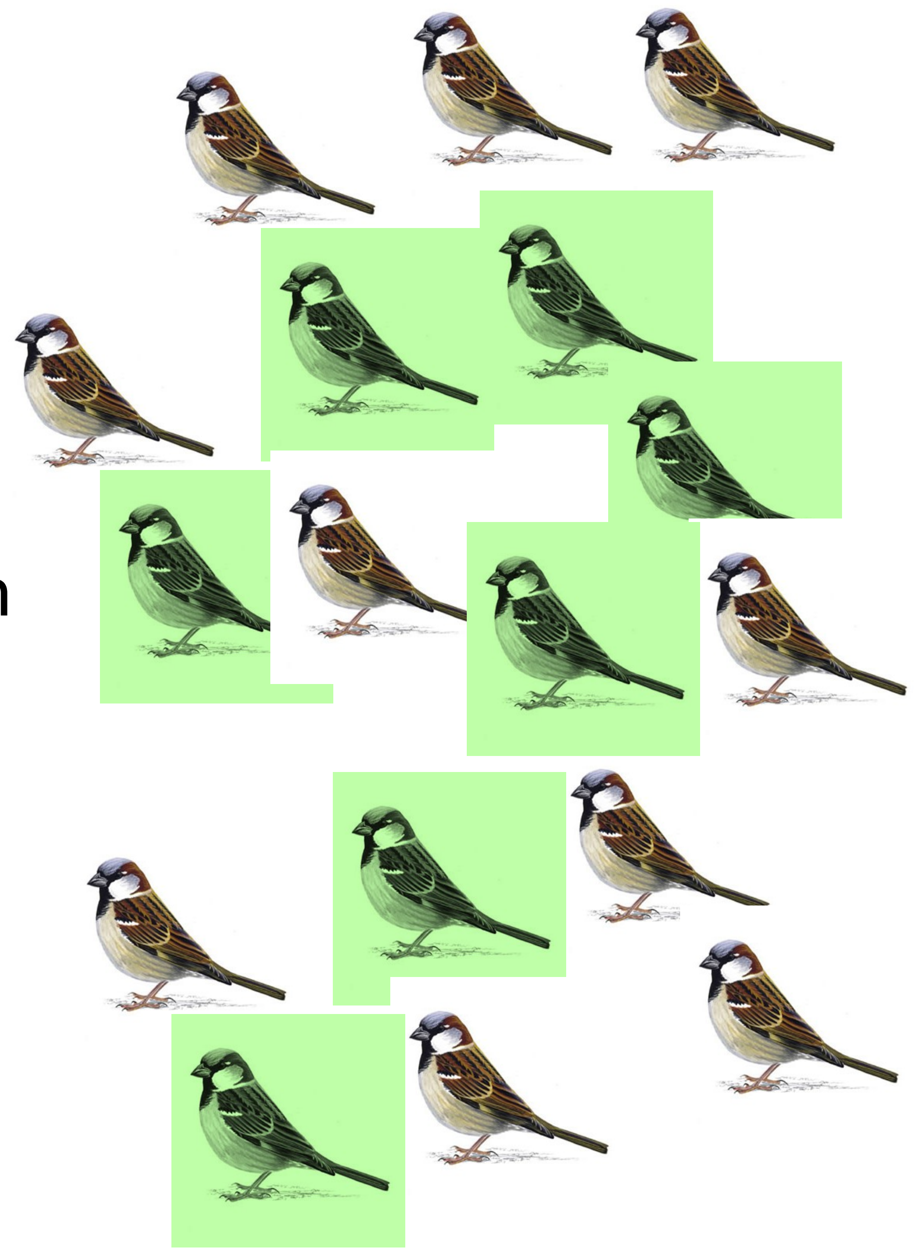


Covariance

- How two variables change together

- Population: joint probability distribution

- Sample: covariance estimate



Covariance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

$$Cov_{x,y} = \frac{\sum(x-\overline{x})\sum(y-\overline{y})}{n-1}$$

- https://i.imgur.com/cWwxYa9.gifv

Covariance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

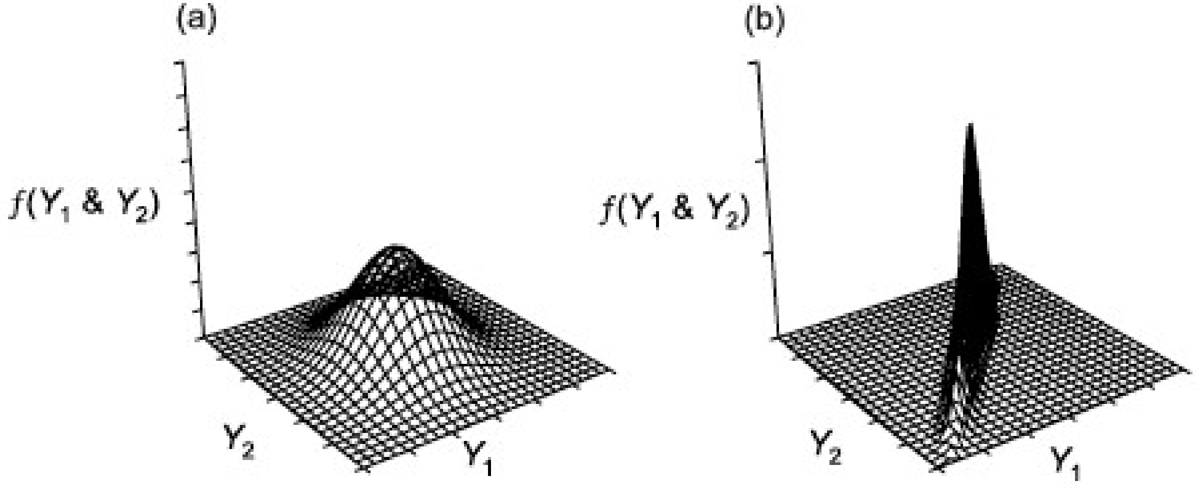
$$Cov_{x,y} = \frac{\sum(x-\overline{x})\sum(y-\overline{y})}{n-1}$$

$$Cov_{x,y} = \frac{\sum(xy) - n\,\overline{x}\,\overline{y}}{n-1}$$

Covariance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

$$Cov_{x,y} = \frac{\sum (x - \overline{x}) \sum (y - \overline{y})}{n-1}$$



Covariance vs Correlation

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

$$Cov_{x,y} = \frac{\sum (x - \overline{x}) \sum (y - \overline{y})}{n-1}$$

$$Cor\,\rho_{x,y} = \frac{Cov_{x,y}}{\sigma_x \sigma_y}$$

Covariance vs Correlation

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

- Ok, why do we need two versions of this?

$$Cov_{x,y} = \frac{\sum(x-\overline{x})\sum(y-\overline{y})}{n-1}$$

$$Cor\ \rho_{x,y} = \frac{Cov_{x,y}}{\sigma_x \sigma_y}$$

The correlation coefficient is the covariance divided by the product of the standard deviations

It is the *standardized* version of the covariance

Covariance vs Correlation

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Ok, why do we need two versions of this?

$$Cov_{x,y} = \frac{\sum(x-\bar{x})\sum(y-\bar{y})}{n-1}$$

$$Cor\,\rho_{x,y} = \frac{Cov_{x,y}}{\sigma_x \sigma_y}$$

The correlation coefficient is the covariance divided by the product of the standard deviations

It is the *standardized* version of the covariance

$r_{x,y}$ =Pearson's correlation coefficient

Correlation coefficient r

Coefficient of determination R^2

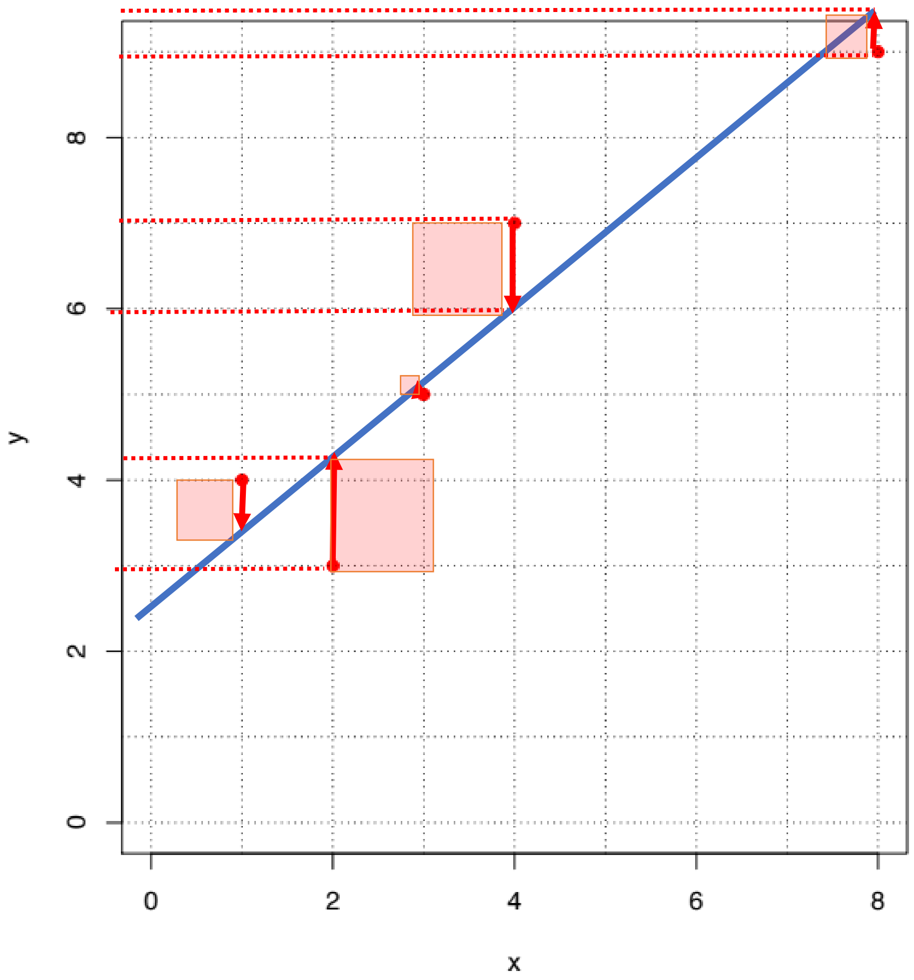
- Describes the relationship between x an y
- Between -1 and 1

- Describes how strong x and y are correlated
- Between 0 and 1

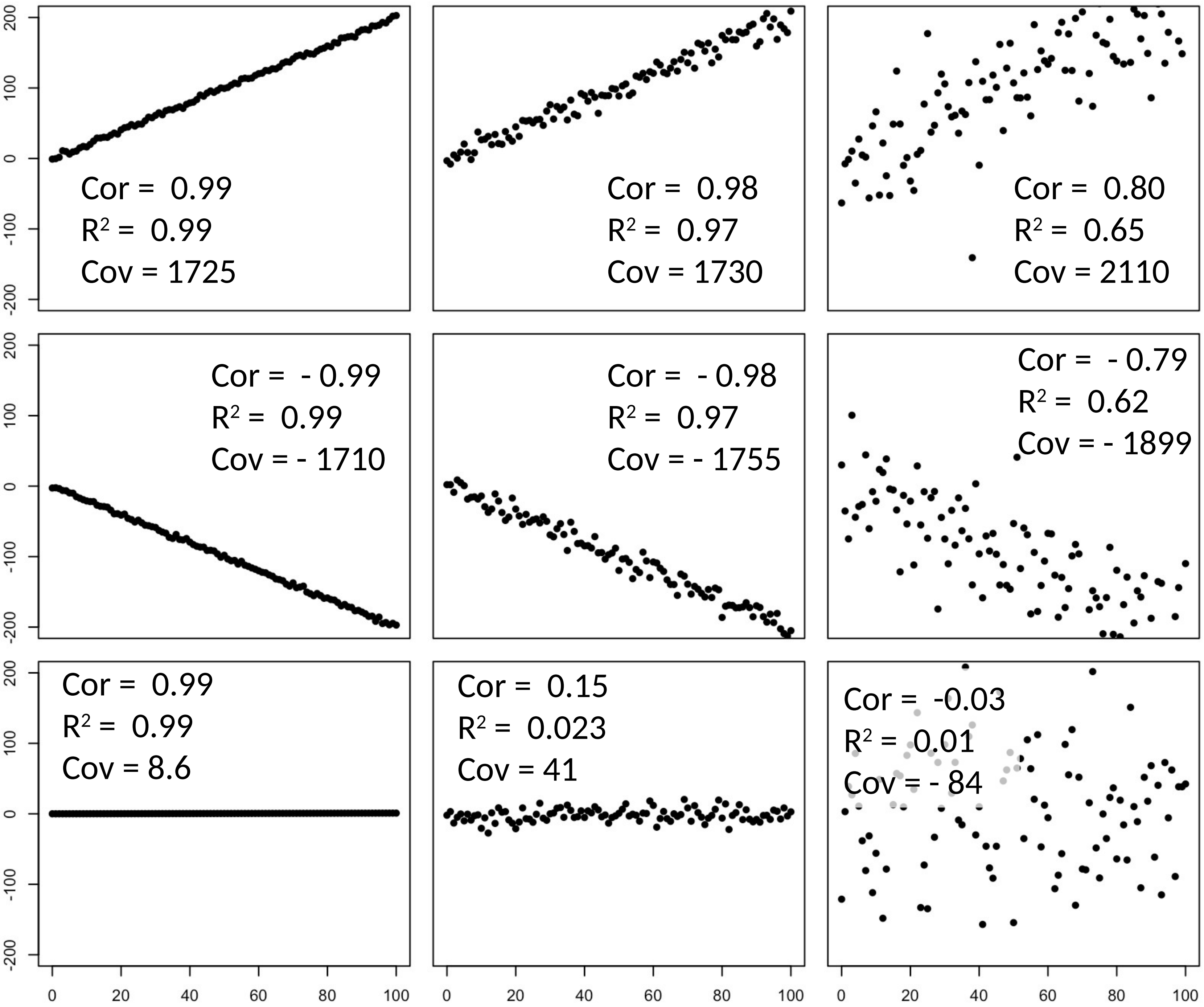
R^2

- Coefficient of determination
- Proportion of how much variance in y is explained by x
- One explanatory variable: r^2

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$



- Or: How much variance remains unexplained



Cor = 0.99
R2 = 0.99
Cov = 1725
Cor = 0.98
R2 = 0.97
Cov = 1730
Cor = 0.80
R2 = 0.65
Cov = 2110
Cor = - 0.99
R2 = 0.99
Cov = - 1710
Cor = - 0.98
R2 = 0.97
Cov = - 1755
Cor = - 0.79
R2 = 0.62
Cov = - 1899
Cor = 0.99
R2 = 0.99
Cov = 8.6
Cor = 0.15
R2 = 0.023
Cov = 41
Cor = -0.03
R2 = 0.01
Cov = - 84
200
100
0
-100
-200
0
20
40
60
80
100

Take home:

- Differences between correlation coefficient, R^2, slope, and covariance