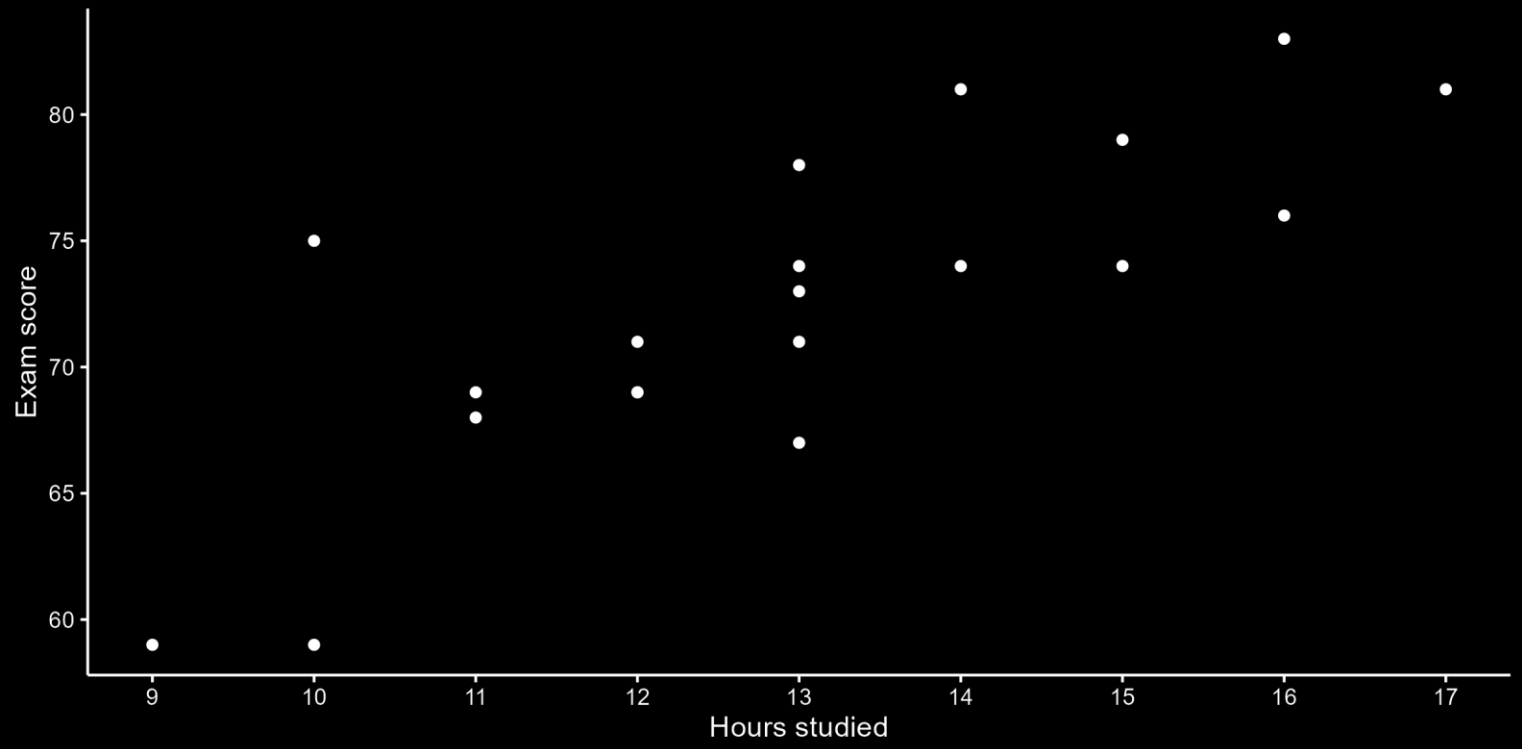
Modeling continuous relationships: Correlation

NHH
TECH3

Example:
Exam scores vs time spent studying



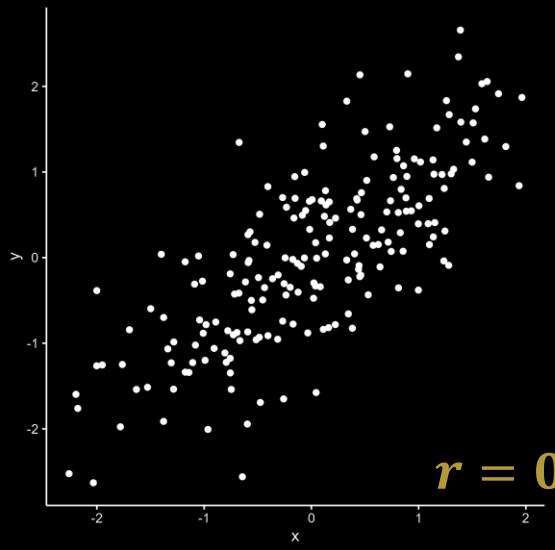
NHH
TECH3

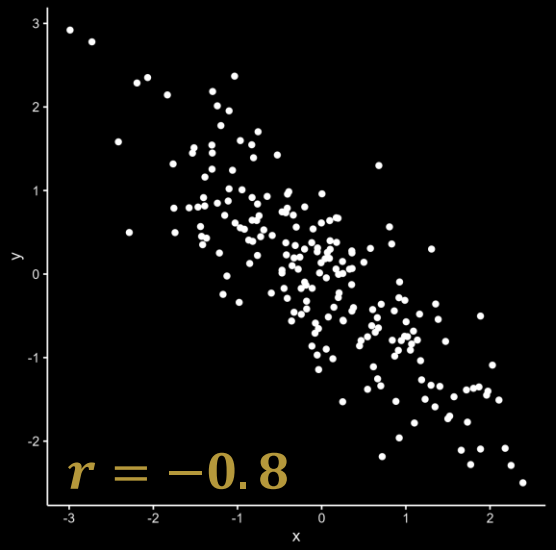
Covariance

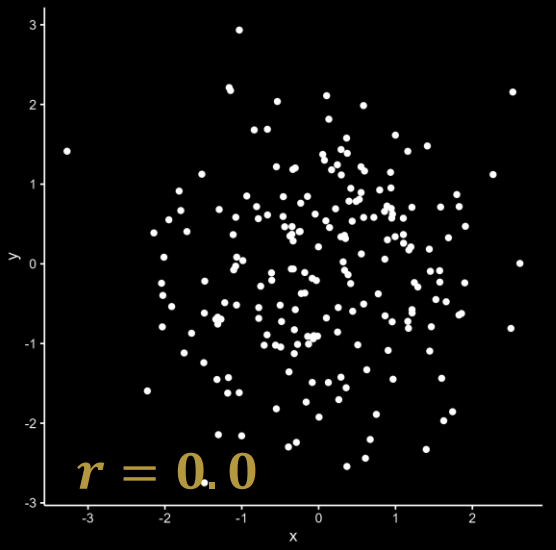
- Variance for a single variable is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- $Covariance = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(\mathrm{y_i} - \bar{y})$
- Correlation coefficient

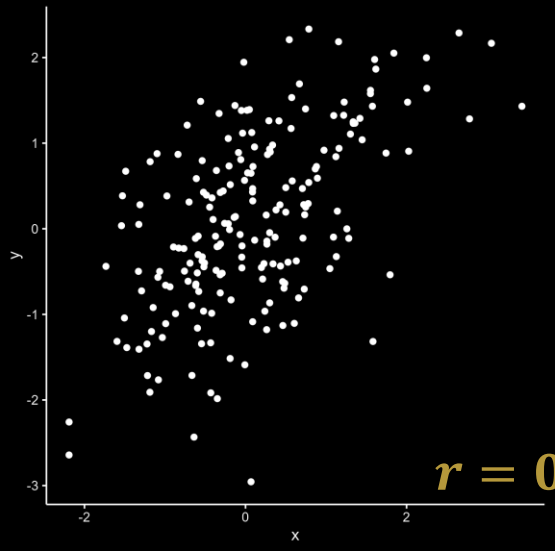
$$r = \frac{Covariance}{s_x\, s_y} = \frac{1}{(n-1)\, s_x\, s_y}\sum_{i=1}^{n}(x_i - \bar{x})(\mathrm{y_i} - \bar{y})$$

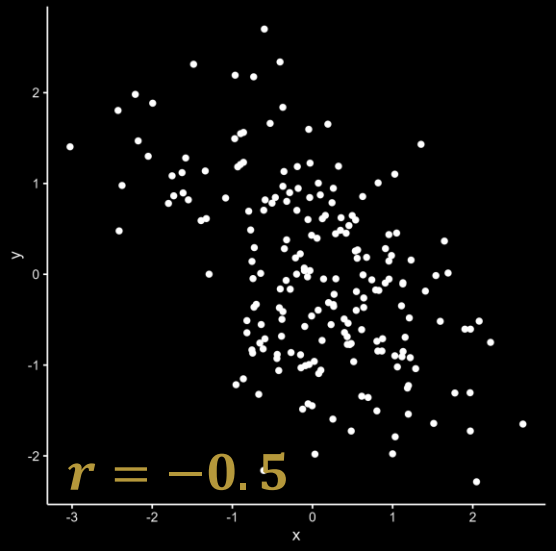
NHH
TECH3





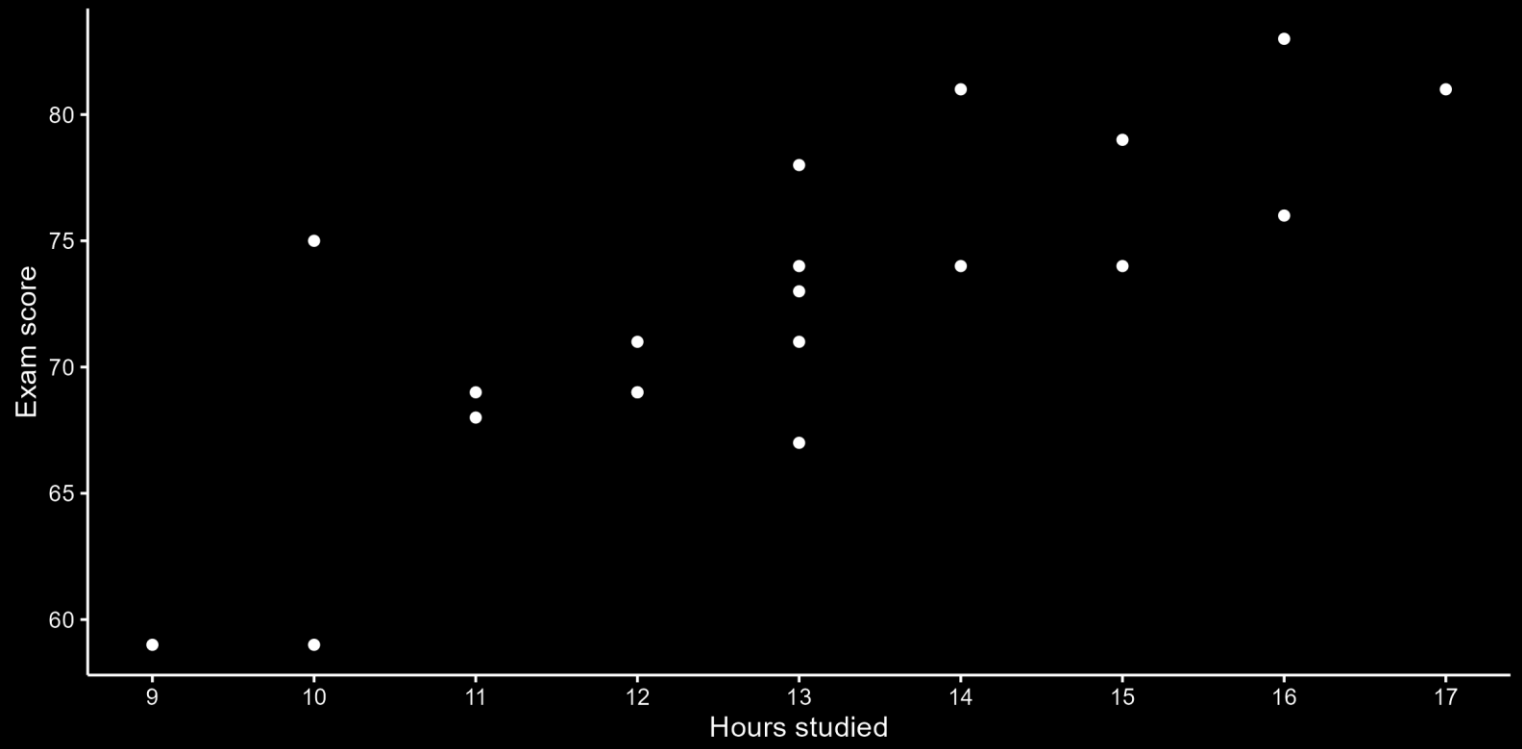






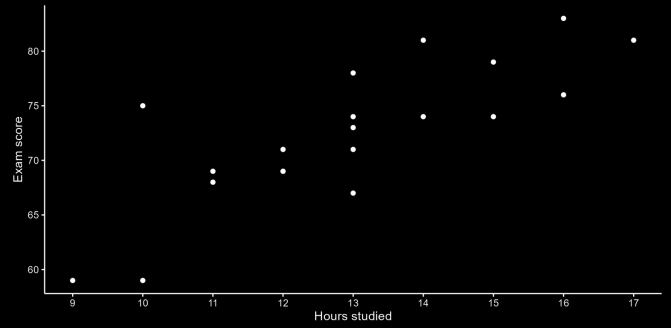
NHH
TECH3

Example:
Exam scores vs time spent studying



NHH
TECH3

Example:
Exam scores vs time spent studying



$$Covariance = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(\mathrm{y_i} - \bar{y}) = 11.2$$

$$r = \frac{Covariance}{s_x\, s_y} = \frac{11.2}{2.16 \cdot 6.53} = 0.796$$

NHH
TECH3

Hypothesis testing for correlations

$H_0: r = 0$

The test statistic

$$T_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

follows a t-distribution with n-2 degrees of freedom under the null hypothesis

Note: This test assumes both variables (X and Y) are normally distributed!

NHH
TECH3

Hypothesis testing for correlations

$H_0: r = 0$ vs $H_A: r > 0$

The test statistic

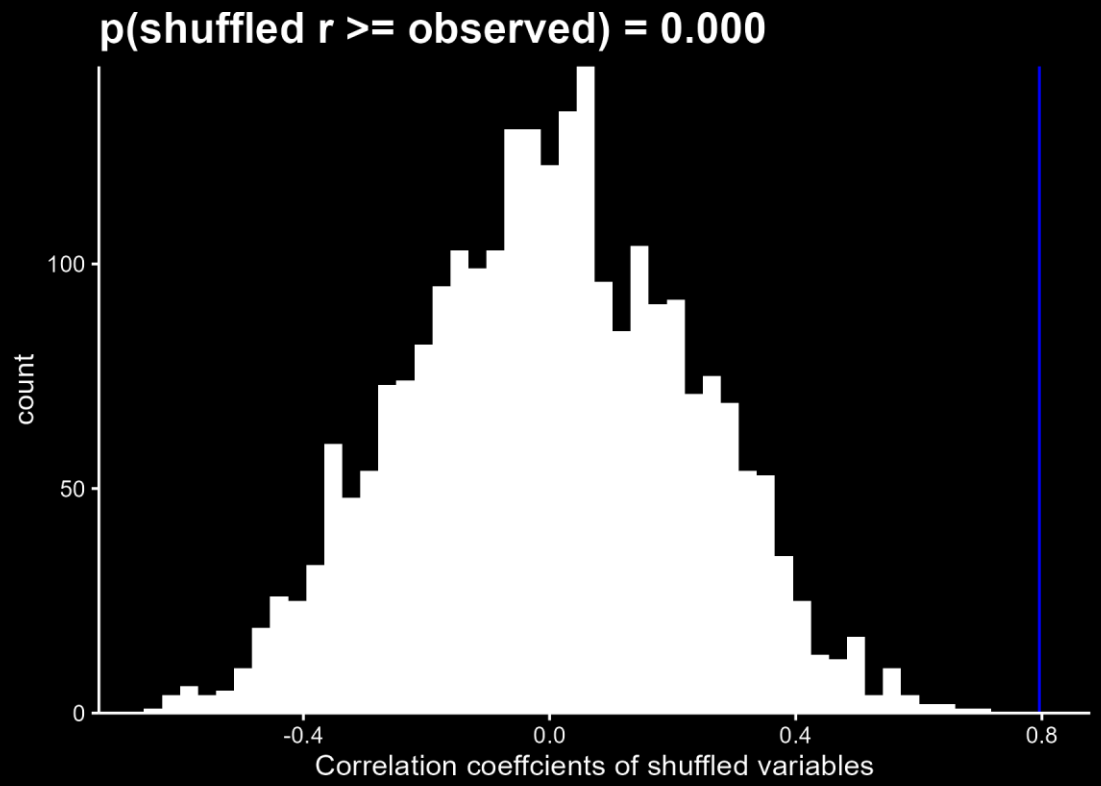
$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.796\sqrt{20-2}}{\sqrt{1-0.796^2}} = 5.58$$

follows a t-distribution with n-2 degrees of freedom under the null hypothesis

```
from scipy import stats
print("p-value: ",
      1-stats.t.cdf(5.58, df = 18))
p-value:  1.3468159696605753e-05
```

NHH
TECH3

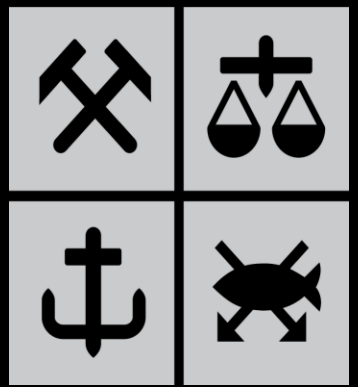
Correlation test by randomization



NHH
TECH3

NHH

TECH3



Sondre Hølleland
Geir Drage Berentsen