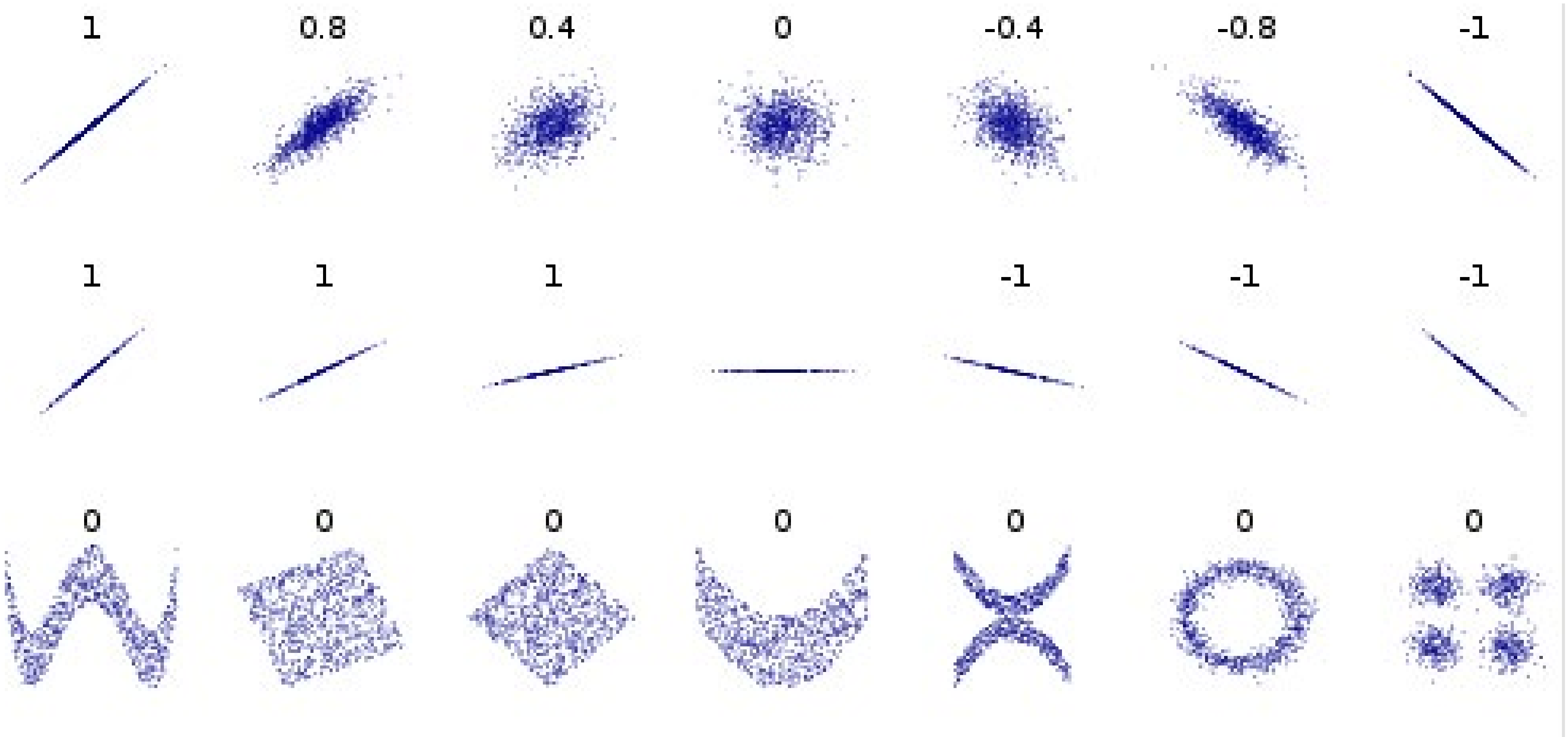
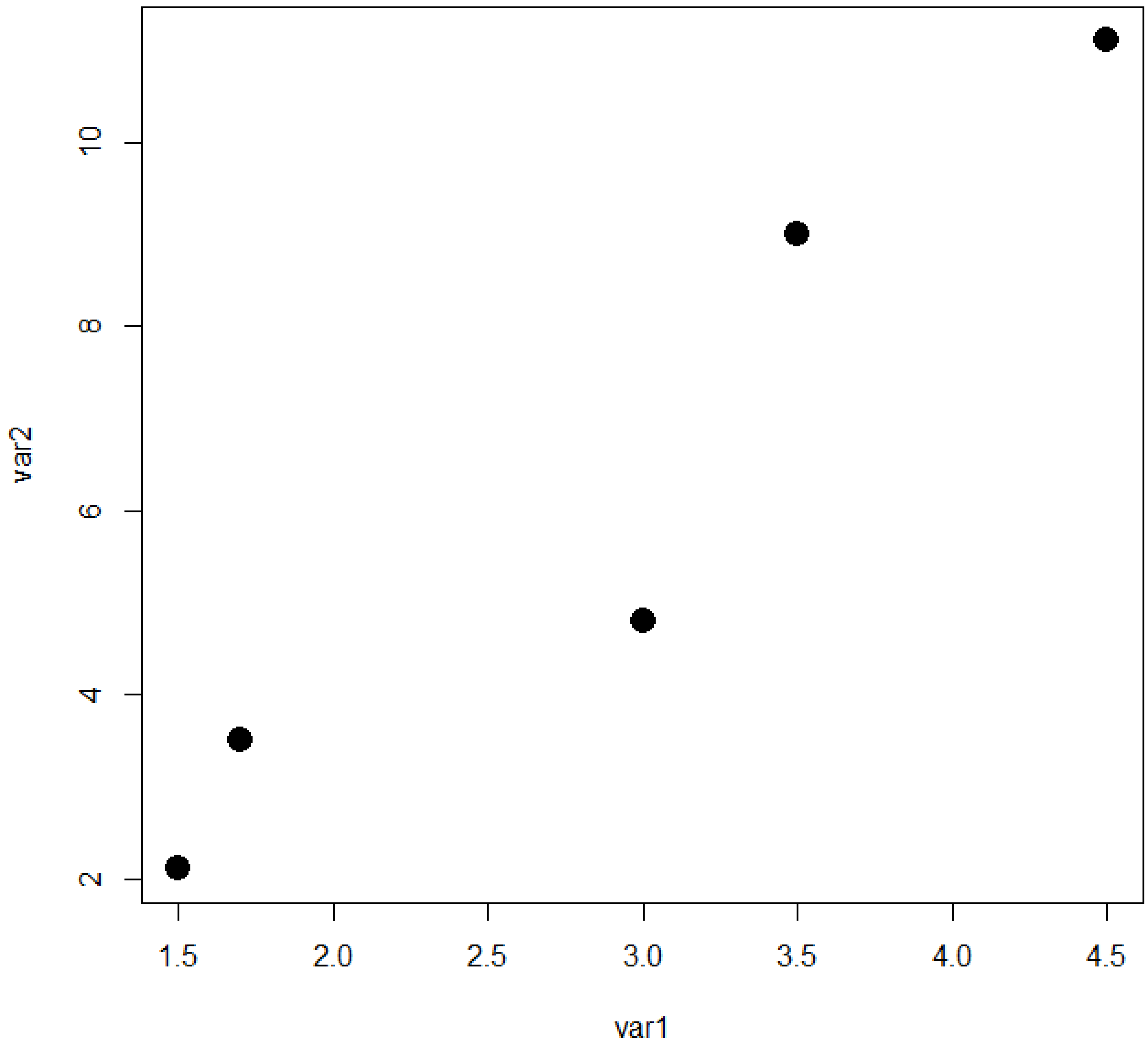
# Correlation and Covariance




Pearson's Correlation ***r*** for various bivariate scatter plots (Source: Wikipedia).

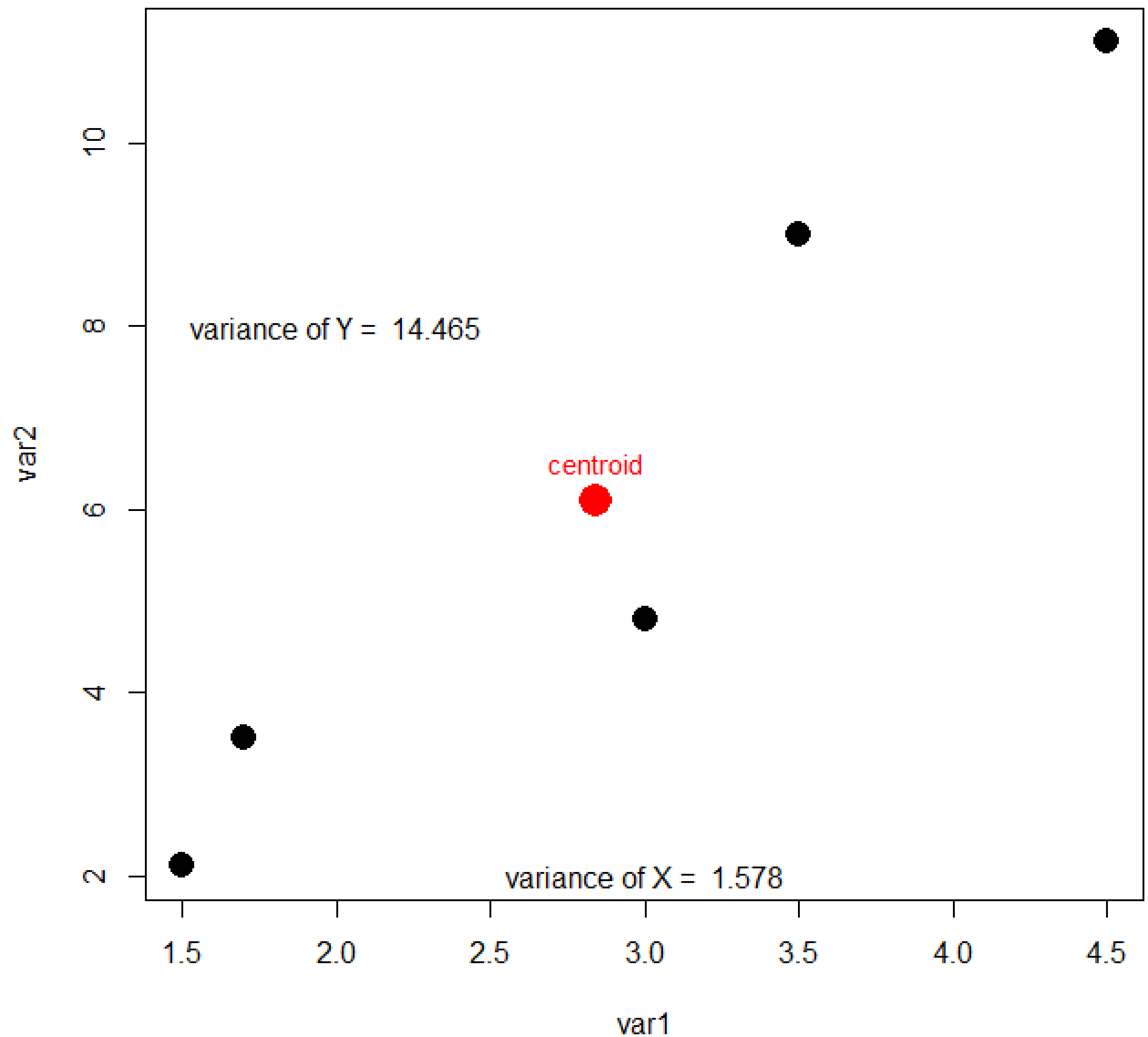
Let's consider a simple bivariate dataset with 5 observations described by two continuous variables X and Y

```
       X    Y
[1,] 1.5  2.1
[2,] 1.7  3.5
[3,] 3.0  4.8
[4,] 3.5  9.0
[5,] 4.5 11.1
> 
```



Note that variables differ notably in variance (Y is much more variable than X)
Red dot marks a ‘centroid’: a bivariate arithmetic mean defined by mean X and mean Y

```
        X    Y
[1,] 1.5  2.1
[2,] 1.7  3.5
[3,] 3.0  4.8
[4,] 3.5  9.0
[5,] 4.5 11.1
> 
```

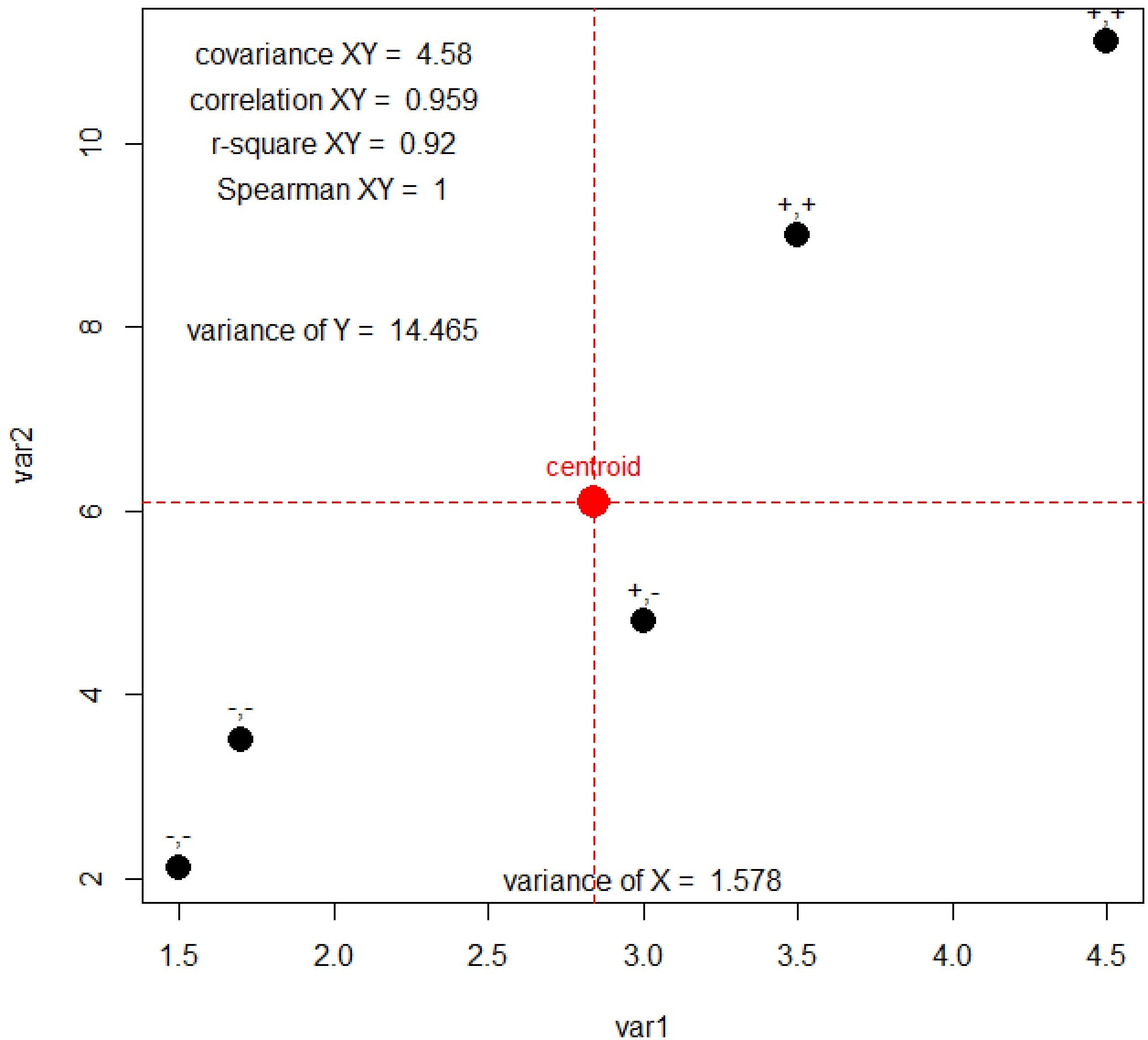


Covariance and Pearson’s correlation both measure the strength and direction (positive or negative) of interrelations of X and Y

R-square is discussed in the subsequent lecture about regression

Spearman rank correlation is a rank-based measure of association.

```
       X    Y
[1,] 1.5  2.1
[2,] 1.7  3.5
[3,] 3.0  4.8
[4,] 3.5  9.0
[5,] 4.5 11.1
> 
```



EXERCISE: Let’s compute covariance and Pearson correlation by hand

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

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

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)}}$$

## Variance

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1} = \frac{\left(\vec{x} - \bar{x}\right)' \cdot \left(\vec{x} - \bar{x}\right)}{n-1}$$

## Covariance

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n-1} = \frac{(\vec{x} - \bar{x})' \cdot (\vec{y} - \bar{y})}{n-1}$$

**Pearson's Correlation – Covariance standardized for variance (covariance divided by product of standard deviations)**

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{s_x s_y} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

## Summary I

Variance (one variable x) – Sum of Squares/(n-1)

Covariance (for two variables: x, y) – Sum of products of deviations of x and y

Is magnitude of covariance independent from magnitude of variance? NO

What is the possible range of values for covariance? $-\infty$ to $\infty$

Is magnitude of correlation independent from magnitude of variance? YES

What is the possible range of values for Pearson's correlation? -1 to 1

# Spearman Rank Correlation

## Pearson

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

| | |
|---|---|
| x | y |
| 2.87 | 0.94 |
| 2.32 | 1.46 |
| 0.5 | 27.5 |
| 0.4 | 250 |

sd(x): 1.259269 sd(y): 120.6555

cov: -102.143

r = -102.143 / (1.259269* 120.6555)

r = -0.6720137

## Spearman

$$r_{xy} = \frac{\mathrm{cov}(r_x, r_y)}{s_{rx} s_{ry}}$$

| | |
|---|---|
| x | y |
| 4 | 1 |
| 3 | 2 |
| 2 | 3 |
| 1 | 4 |

sd: 1.290994 sd: 1.290994

cov: -1.666667

r = -1.666667 / (1.290994*1.290994)

r = -1

# Spearman Rank Correlation

## Pearson

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

| x | y |
|---|---|
| 2.32 | 0.94 |
| 2.87 | 1.46 |
| 0.5 | 27.5 |
| 0.4 | 250 |

sd(x): 1.259269 sd(y): 120.6555

cov: -102.0089 (-102.143)

r = -102.0089 / (1.259269* 120.6555)

r = -0.6713862 (-0.6720137)

## Spearman

$$r_{xy} = \frac{\mathrm{cov}(r_x, r_y)}{s_{r_x} s_{r_y}}$$

| x | y |
|---|---|
| 3 | 1 |
| 4 | 2 |
| 2 | 3 |
| 1 | 4 |

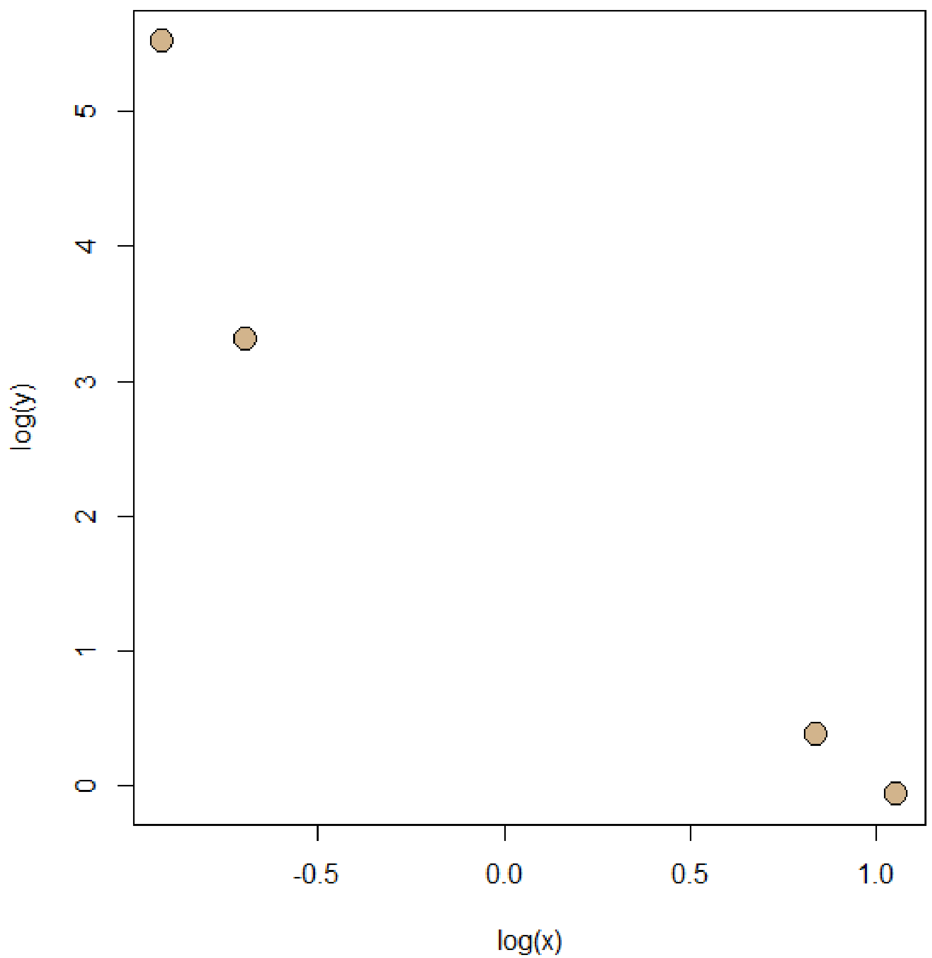
sd: 1.290994 sd: 1.290994

cov: -1.333333 (-1.666667)

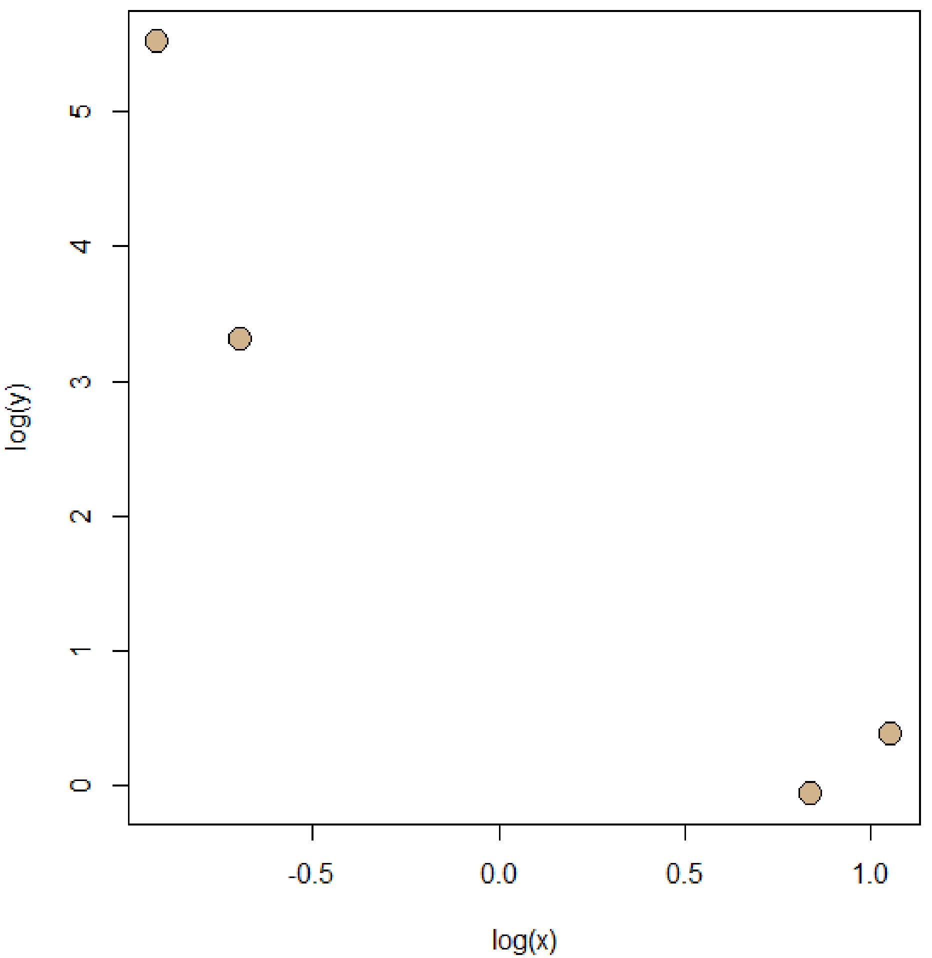
r = -1.333333 / (1.290994*1.290994)

r = -0.8 (-1)

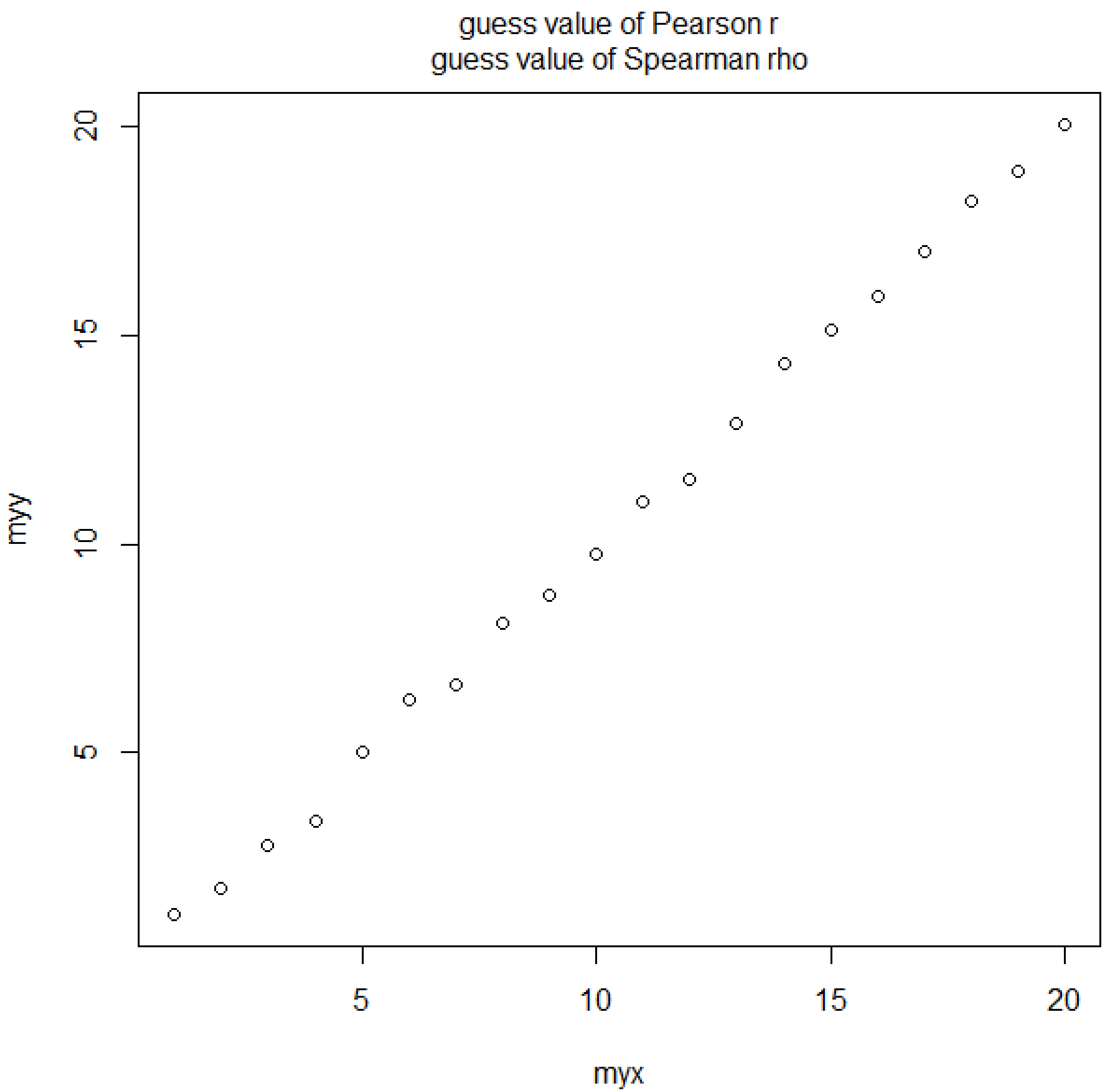
Pearson r = - 0.9660809
Spearman = 1



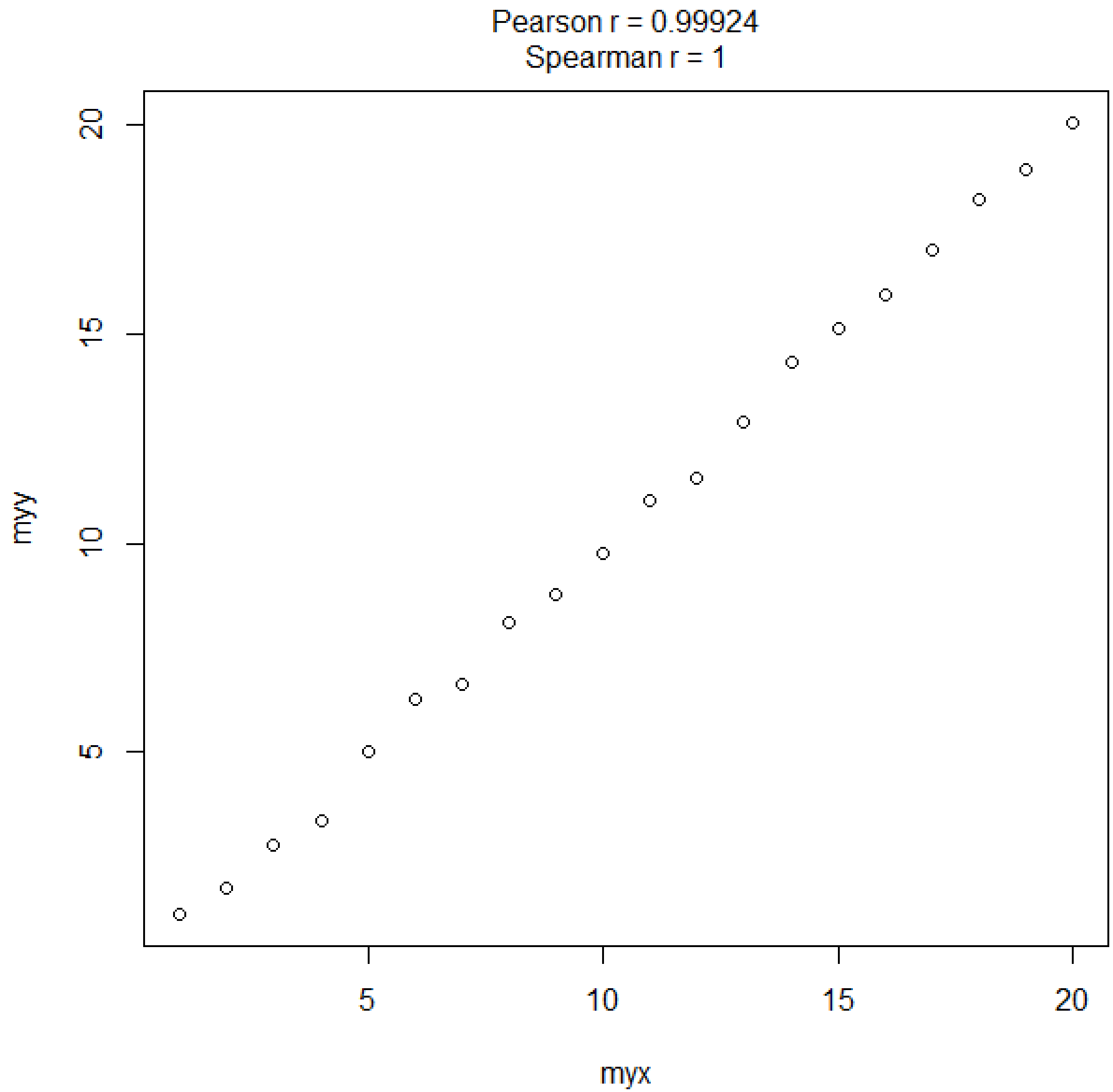

Pearson r = - 0.9544169
Spearman = 0.8



# Spearman versus Pearson

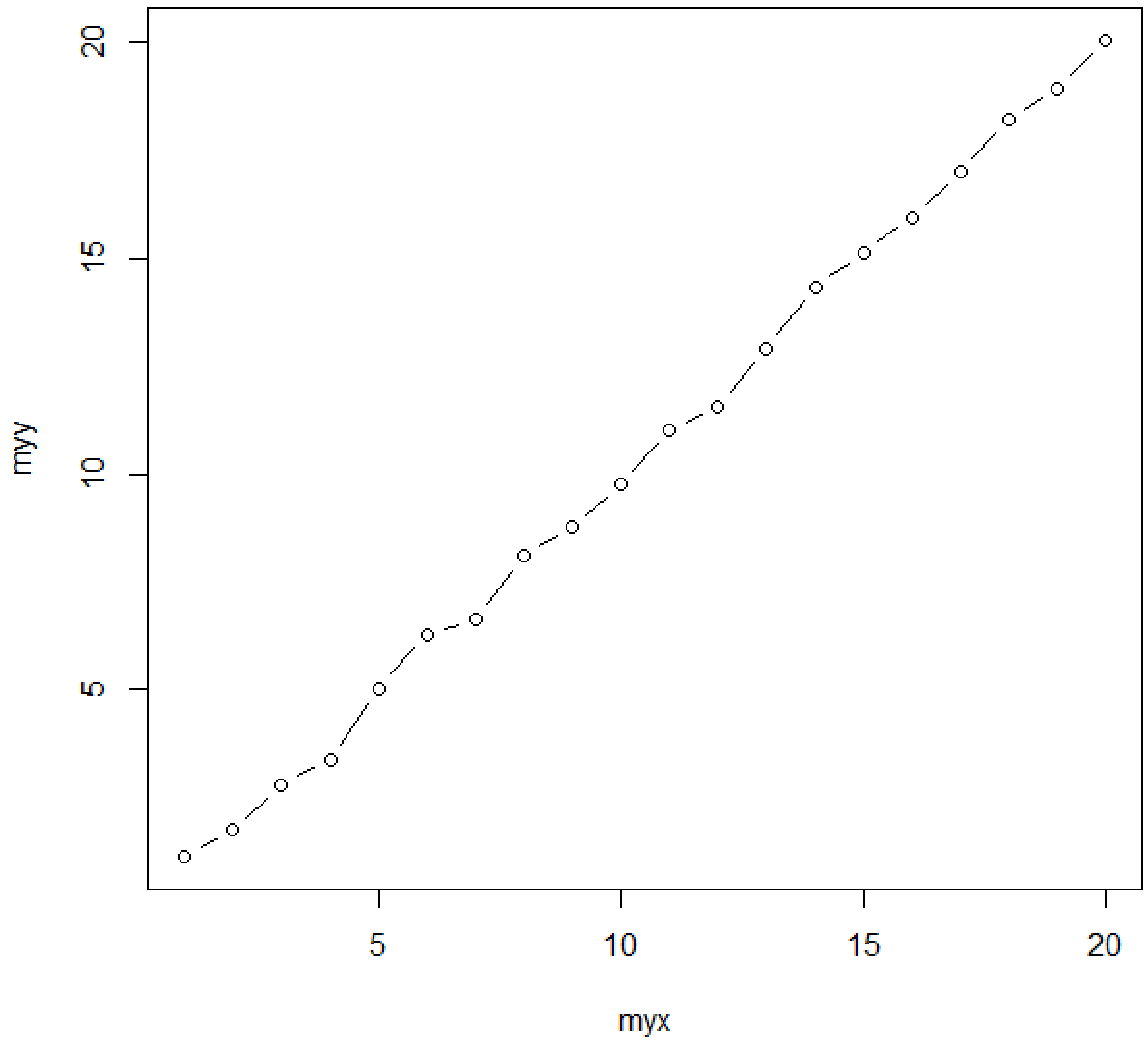


# Spearman versus Pearson

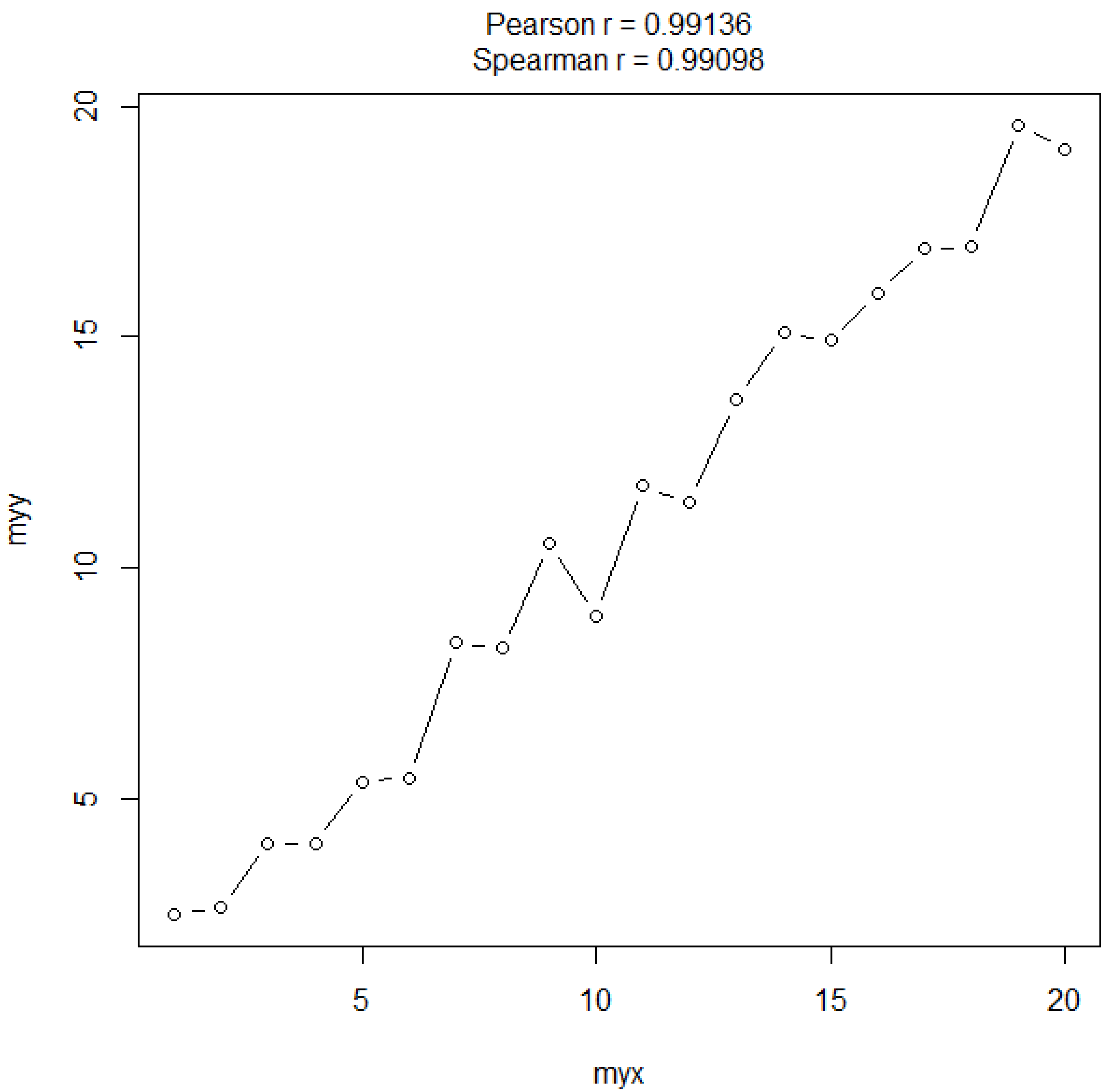


Spearman versus Pearson

Two monotonically related variables will yield Spearman = 1 (or -1)

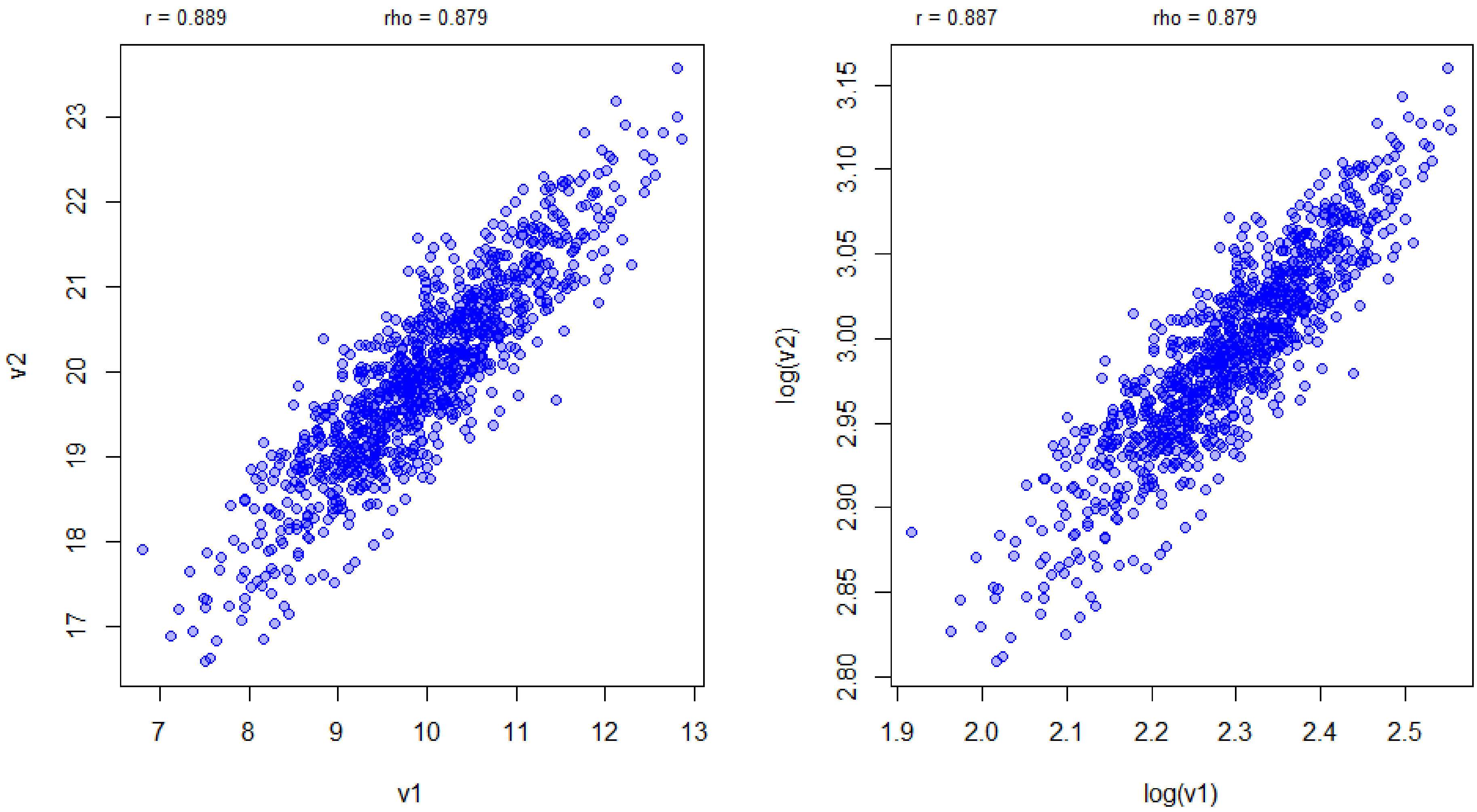


# Spearman versus Pearson



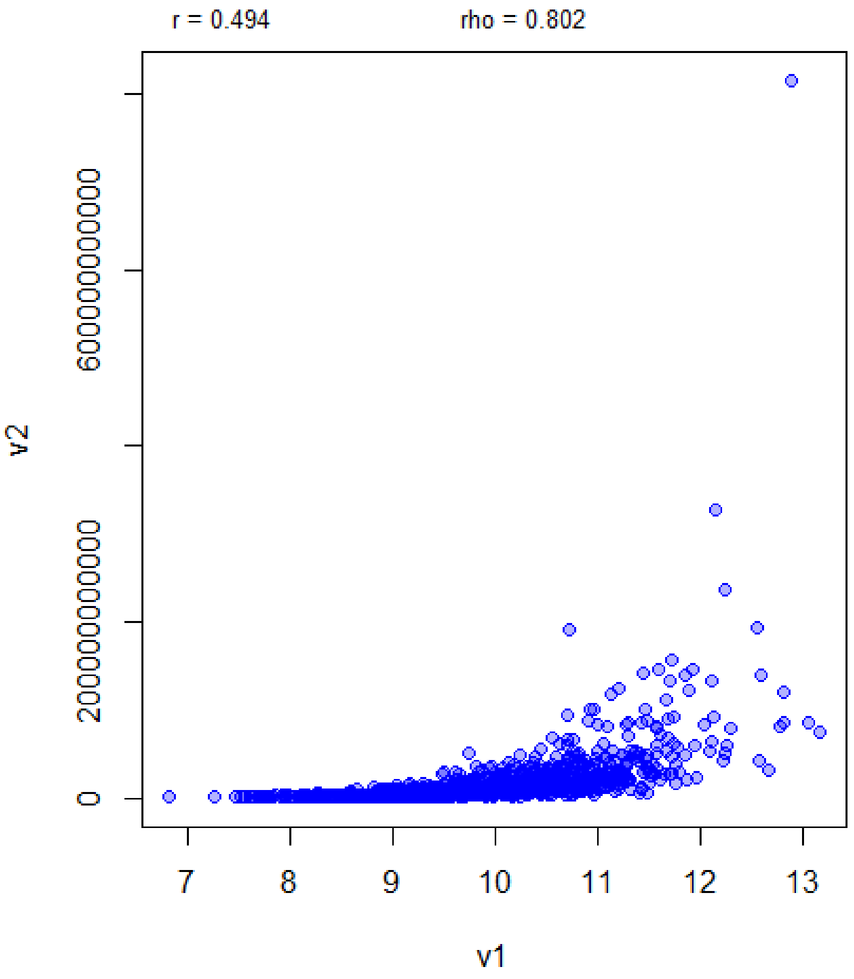
# Spearman versus Pearson

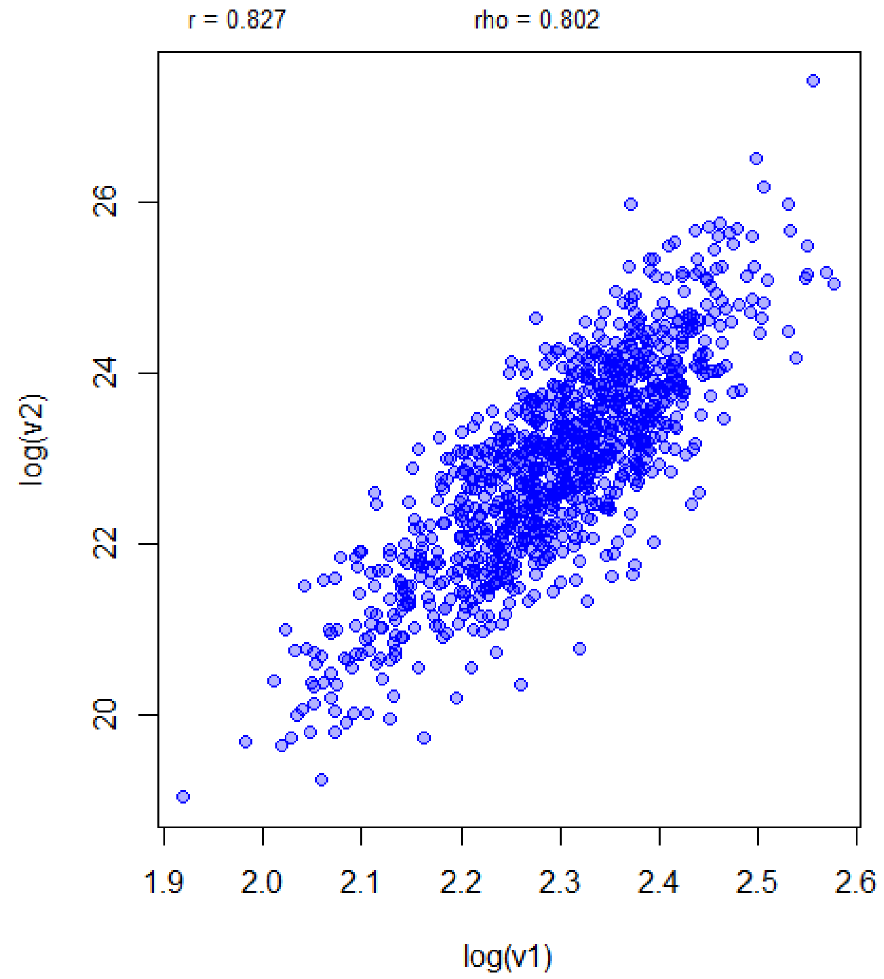
*normally distributed and well-behaved data*



# Spearman versus Pearson

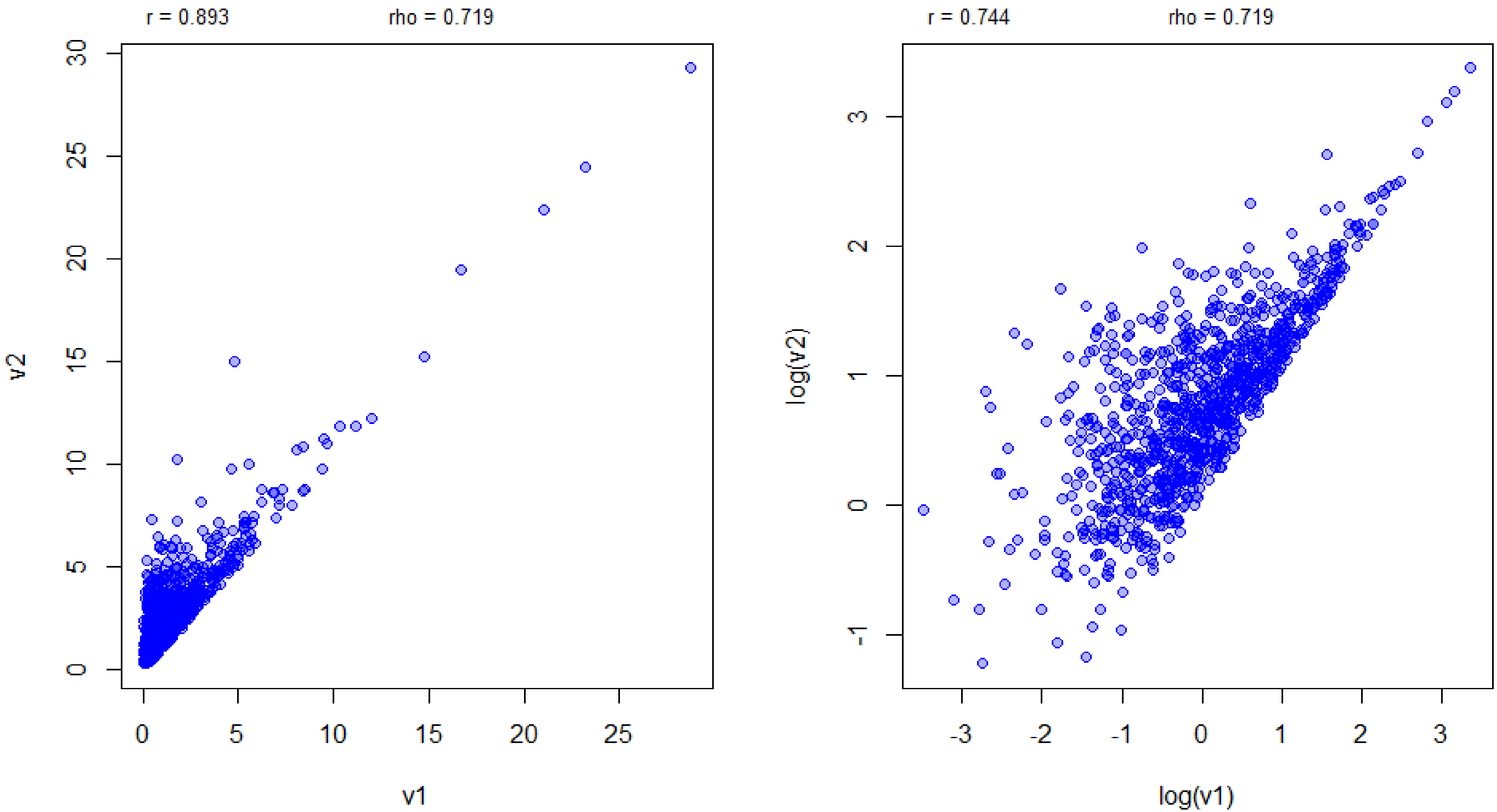
*two non-linearly correlated variables*
*(log-transformation often linearizes the relationship)*



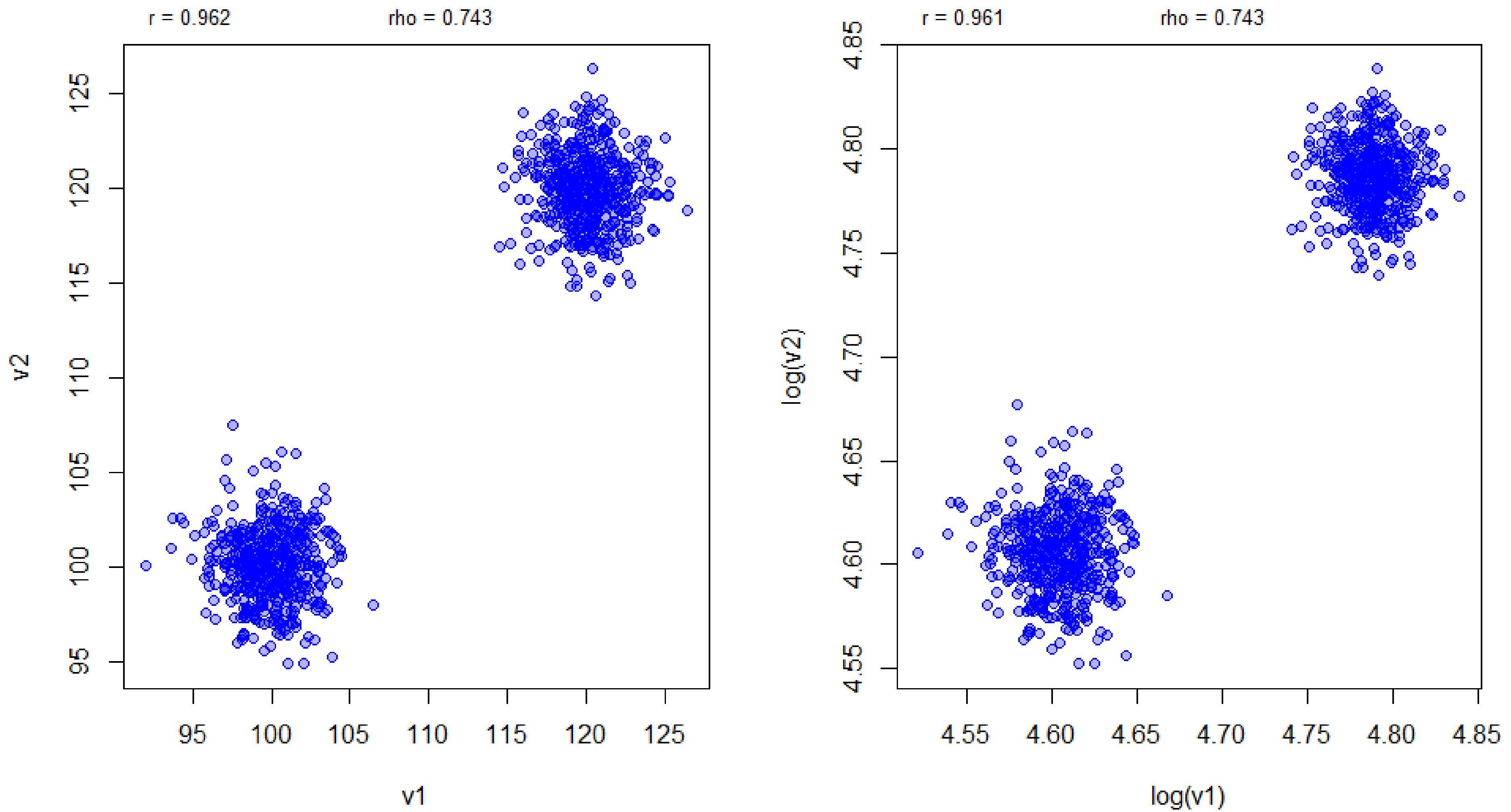
# Spearman versus Pearson

*two correlated variables from non-normal distribution*



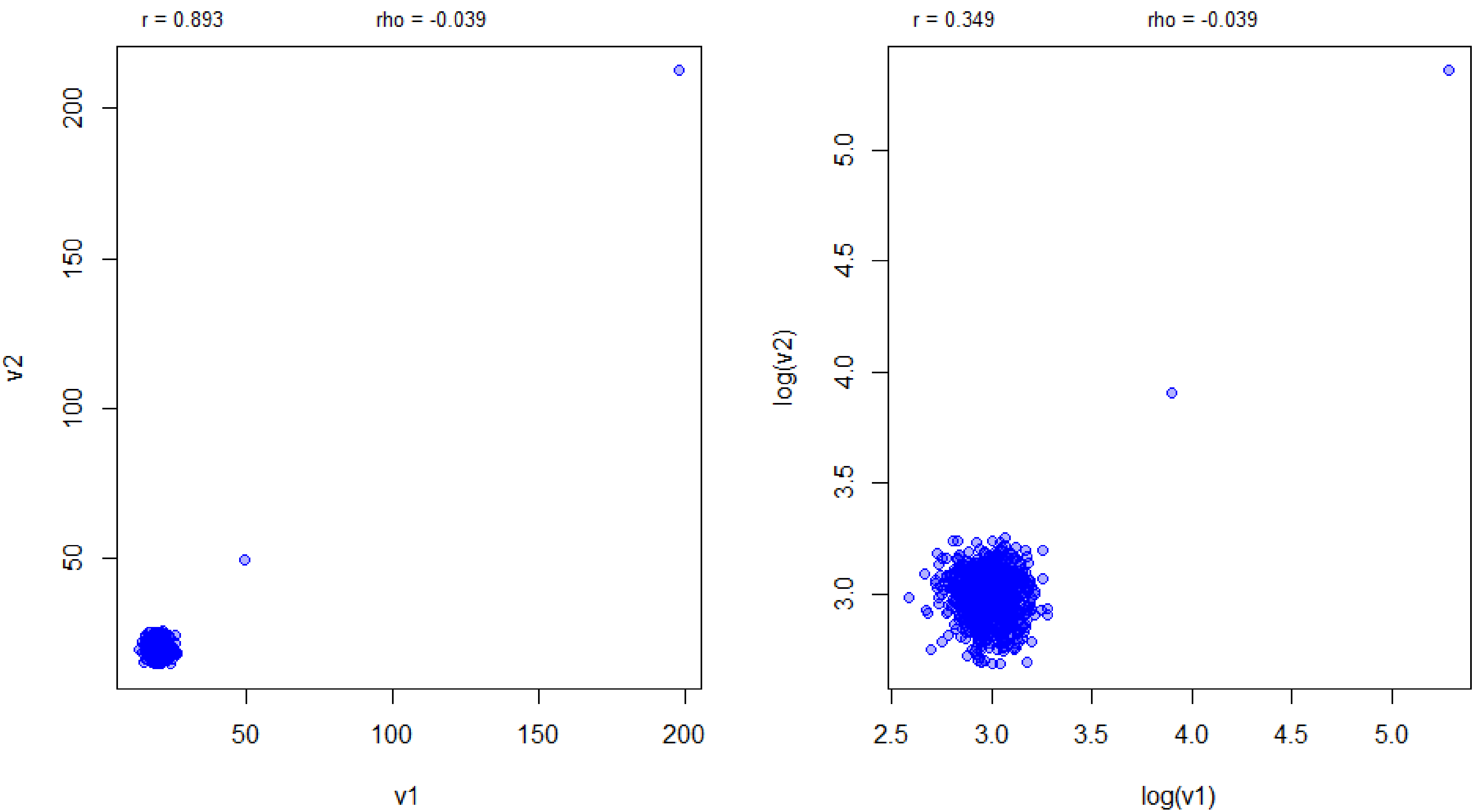
# Spearman versus Pearson

*two variables with bimodal distribution*



# Spearman versus Pearson

*Outliers*
*(rank correlation much more immune to extreme outliers*



# Kendall Rank Correlation

$$\tau = \frac{number\ of\ concordant\ pairs - number\ of\ discordant\ pairs}{0.5 * \mathrm{n} * (\mathrm{n} - 1)}$$

$$\tau_B = \frac{number\ of\ concordant\ pairs - number\ of\ discordant\ pairs}{\sqrt{t_1} * \sqrt{t_2}}$$

Tau-B Is used by {cor} and {cor.test} functions in *R*

$0.5 * \mathrm{n} * (\mathrm{n} - 1)\ - total\ number\ of\ possible\ comparisons$
$\sqrt{t_1} * \sqrt{t_2}\ - total\ number\ of\ possible\ comparisons\ (if\ no\ ties)$

$t_1$ – number of non-ties in x, and $t_2$ number of non-ties in y

## Pearson

| x | y |
|---|---|
| 2.87 | 0.94 |
| 2.32 | 1.46 |
| 0.5 | 27.5 |
| 0.4 | 250 |

sd(x): 1.259269   sd(y): 120.6555

cov: -102.143

r = -102.143 / (1.259269* 120.6555)

r = -0.6720137

## Kendall

| Rank of x | Rank of y |
|---|---|
| 4 | 1 |
| 3 | 2 |
| 2 | 3 |
| 1 | 4 |

| | | |
|---|---|---|
| 4 > 3 | but | 1 < 2 discordant pair |
| 4 > 2 | but | 1 < 3 discordant pair |
| 4 > 1 | but | 1 < 4 discordant pair |
| 3 > 2 | but | 2 < 3 discordant pair |
| 3 > 1 | but | 2 < 4 discordant pair |
| 2 > 1 | but | 3 < 4 discordant pair |

0 concordant and 6 discordant pairs
6 possible comparisons

$$\tau = \tau_B = \frac{0 - 6}{\sqrt{6} * \sqrt{6}} = -1$$

## Summary II

Rank correlation coefficients vary from -1 to 1

Rank correlation must be 1 (or -1) if relation is monotonic

Rank correlation of x and y is the same as rank correlation of log(x) and log(y)
(log transformation is monotonic)

Pearson correlation of x and y is <u>NOT</u> the same as Pearson correlation of log(x) and log(y)

Rank correlations tend to be less sensitive to outliers and non-normality

Rank correlations is more suitable for discrete variables

Neither rank correlation nor Pearson correlation can handle strongly bimodal (or multimodal) data

## Testing for significance of correlation coefficient r

Parametric tests (t-statistic or F-statistic) assume bivariate normal distribution

$H_0$: r = 0
$H_A$: r ≠ 0

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

$df = n - 2$

$p = p(t, df, \text{2-tailed})$

Why are there multiple different equations for *t* ?

Are they merely modified variants of the canonical form?

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

$$t = \frac{x - \mu}{\frac{s_x}{\sqrt{n}}}$$

Which of the two is the more correct function for defining t distribution?
First? Second? Either? Neither?

Why is this equation different from t-test equation for means?

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

$$t = \frac{x - \mu}{\frac{s_x}{\sqrt{n}}}$$

Which of the two is the more correct function for defining t distribution?

**Neither!**

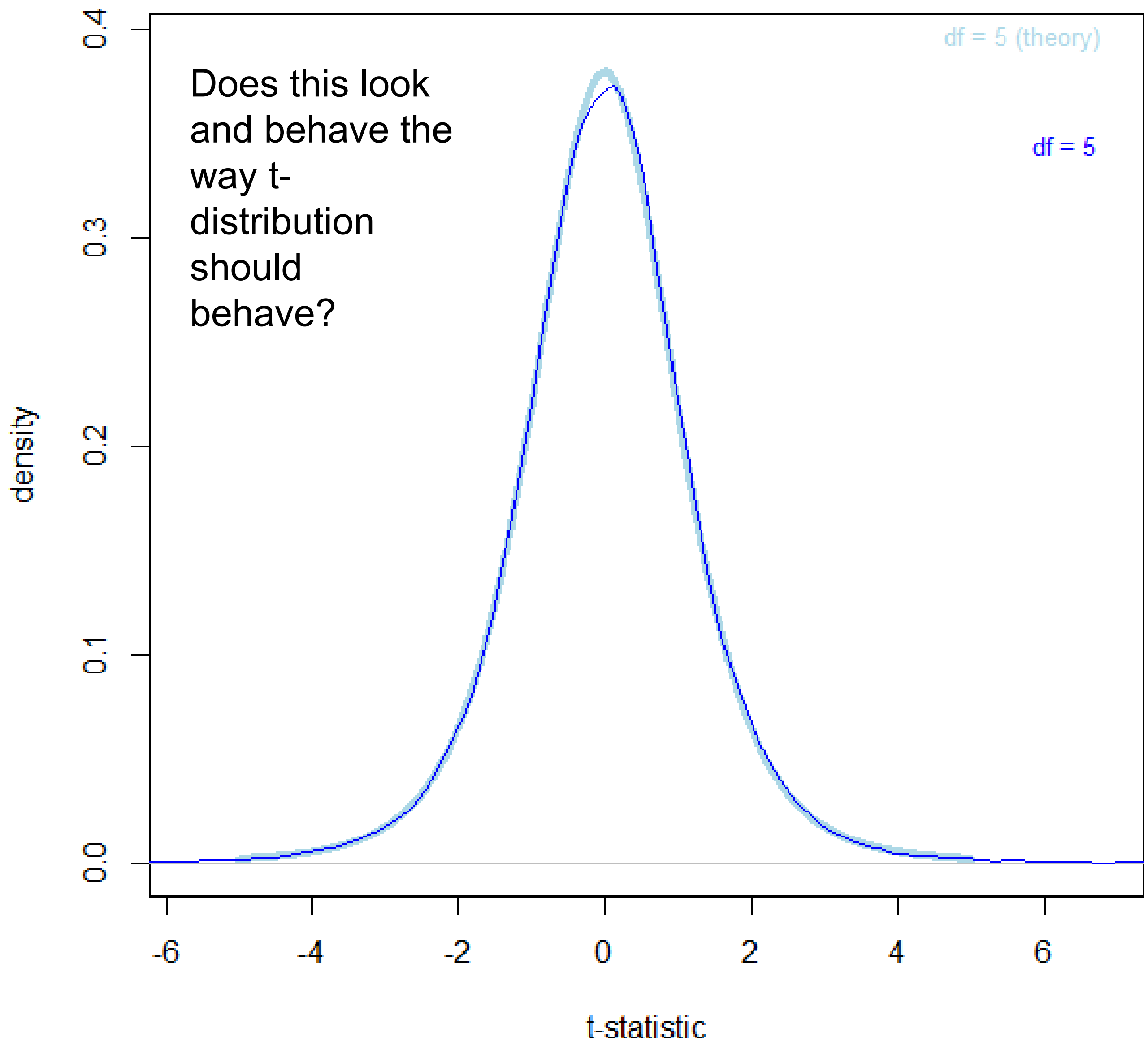
*t* – probability density function

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

Various equations for *t*-statistic are not synonymous with *t* function. These equations produce estimates that are approximately *t* distributed **when assumptions are met**

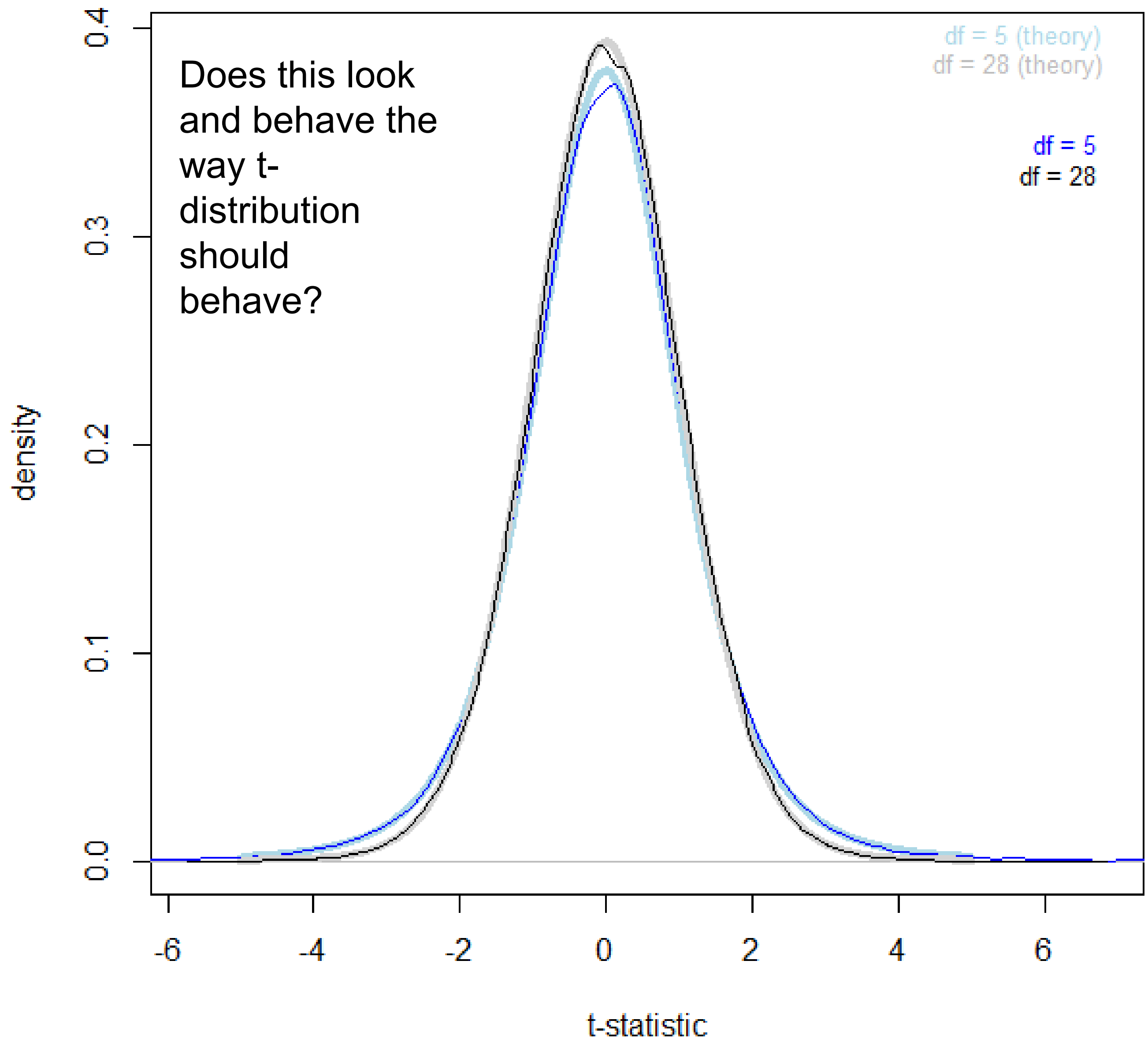
$$t = r * \sqrt{\frac{n-2}{1-r^2}}$$

Simulated distribution of $t$ values for $n$ = 7 for two uncorrelated samples drawn from normal distribution



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

Simulated distribution of $t$ values for $n$ = 7 and $n$ = 30 for two uncorrelated samples drawn from normal distribution



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

Simulated distribution of *t* values for *n* = 7 and *n* = 30 for two uncorrelated samples drawn from normal distribution
+ non-normally distributed data for the same n values.