Solution Sheet 4

The full R code will be made available in a separate file on the course homepage.

1. **a)** See full R code. Using the function `set.seed(N)`, sets the random number generator to state $N \in \mathbb{N}$; this makes the sequence of random numbers reproducible, i.e., the exact same sequence of numbers is generated if `set.seed(N)` is called initially. Here, we used $N = 42$.

b) The correlation matrix of the simulated data can be computed by calling `cor(data)`. In our case, the correlation estimate given by the data is 0.765, which deviates from the theoretical value of 0.7. This is to be expected since our sample is finite ($n = 100$) and thus the correlation estimate (which is a random variable itself) is subject to random fluctuations.

c) We can compute 1000 estimates of correlation (each based on $n = 100$ observations) using the R code given on the exercise sheet. Figure 1 shows a histogram of the resulting estimates. The sample standard deviation of the correlation estimates is 0.049.

d) The larger the number of observations we use, the higher the precision of our correlation estimate becomes. Figure 2 shows the relation between the two.

e) From the theory we know that the standard deviation of the estimator is approximately proportional to $\frac{1}{\sqrt{n}}$. To show that a quantity is proportional to a power (such as x^k) of some other quantity, it is advisable to draw a so-called log-log plot. In such a plot, the logarithms (with respect to any basis) are taken of both x and y values before plotting. The degree to which the power is taken (k if the power is x^k) then appears as the slope in this plot. The following calculation makes this easy to see:

$$\begin{aligned} y &= Cx^k \\ \log(y) &= \log(Cx^k) \\ \log(y) &= \log(C) + k \cdot \log(x) \end{aligned}$$

Figure 3 shows the logarithm of the standard deviation for the correlation estimate plotted against the logarithm of the number of observations. Here a linear relationship is evident. Estimating the slope (using the function `lm` here), we find it to be -0.54. Then, the standard deviation of the correlation estimate is approximately proportional to $n^{-1/2} = \frac{1}{\sqrt{n}}$.

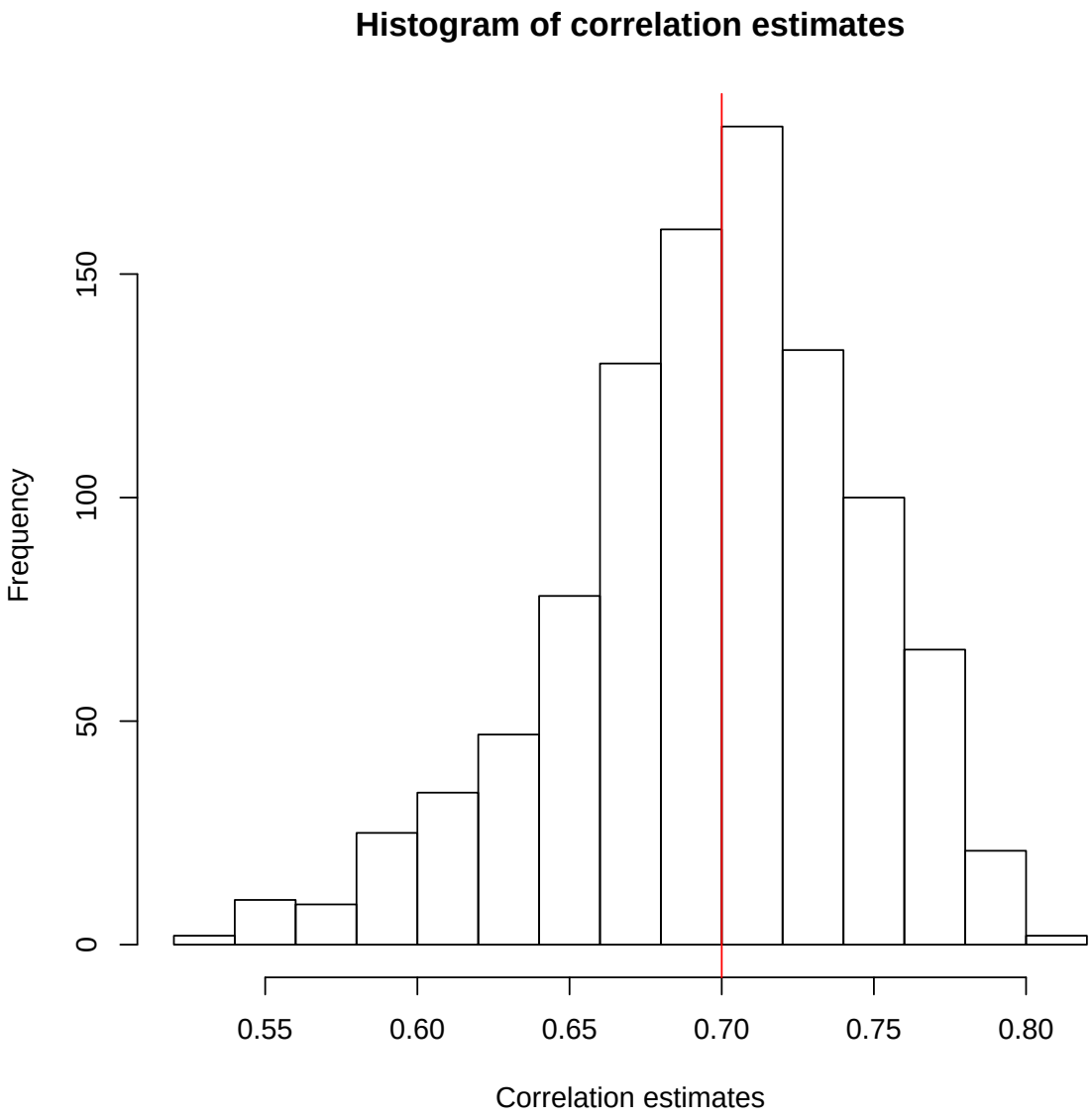

Figure 1: The histogram shows the distribution of 1000 estimates of correlation, each based on $n = 100$ observations. The true correlation is 0.7; this is marked by a solid vertical line.

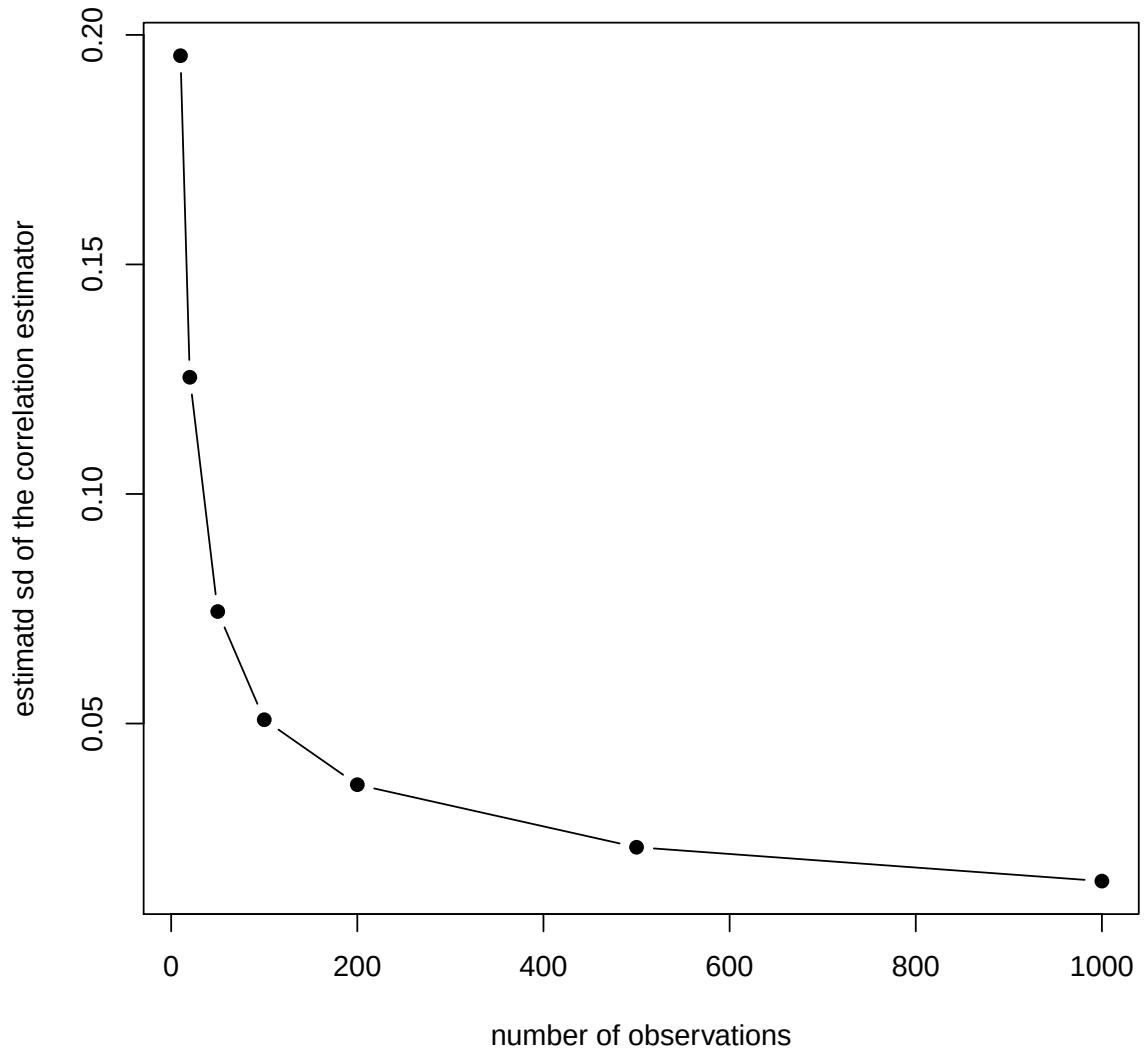

Figure 2: The larger the number of observations used in estimating the correlation, the higher the precision of this estimate. That is, the (estimated) standard deviation of the correlation estimate decreases as n increases.

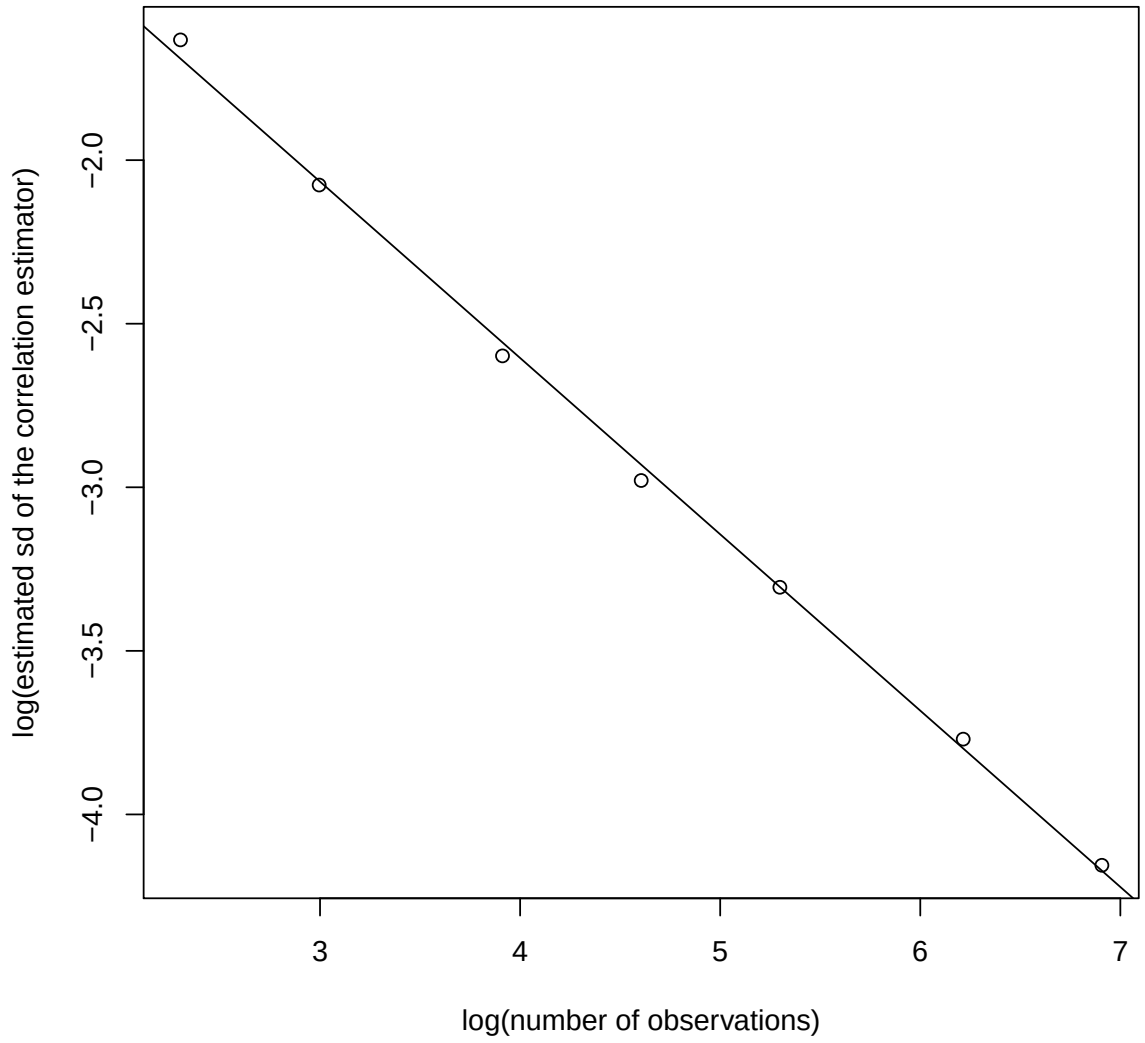

Figure 3: The log-log plot clearly shows a linear relationship with slope about -0.5. That is to say, the standard deviation of the estimator is approximately proportional to $\frac{1}{\sqrt{n}}$.

2. **a)** See full R code.

b) See full R code.

c) See Figure 4. The marginal distributions look normal, but the bulb-shaped scatter plot provides strong evidence against bivariate normality.

d) See Figure 5. The Q-Q plot does not show mentionable evidence against normality, but this does not necessarily mean that there is evidence in favor of normality as the scatter plot in **c)** illustrates.

Remark: The data `clayton.dat` have been simulated from standard normal margins coupled by a so called Clayton copula. Hence, the margins are indeed univariate normal, but the distribution is not bivariate normal.

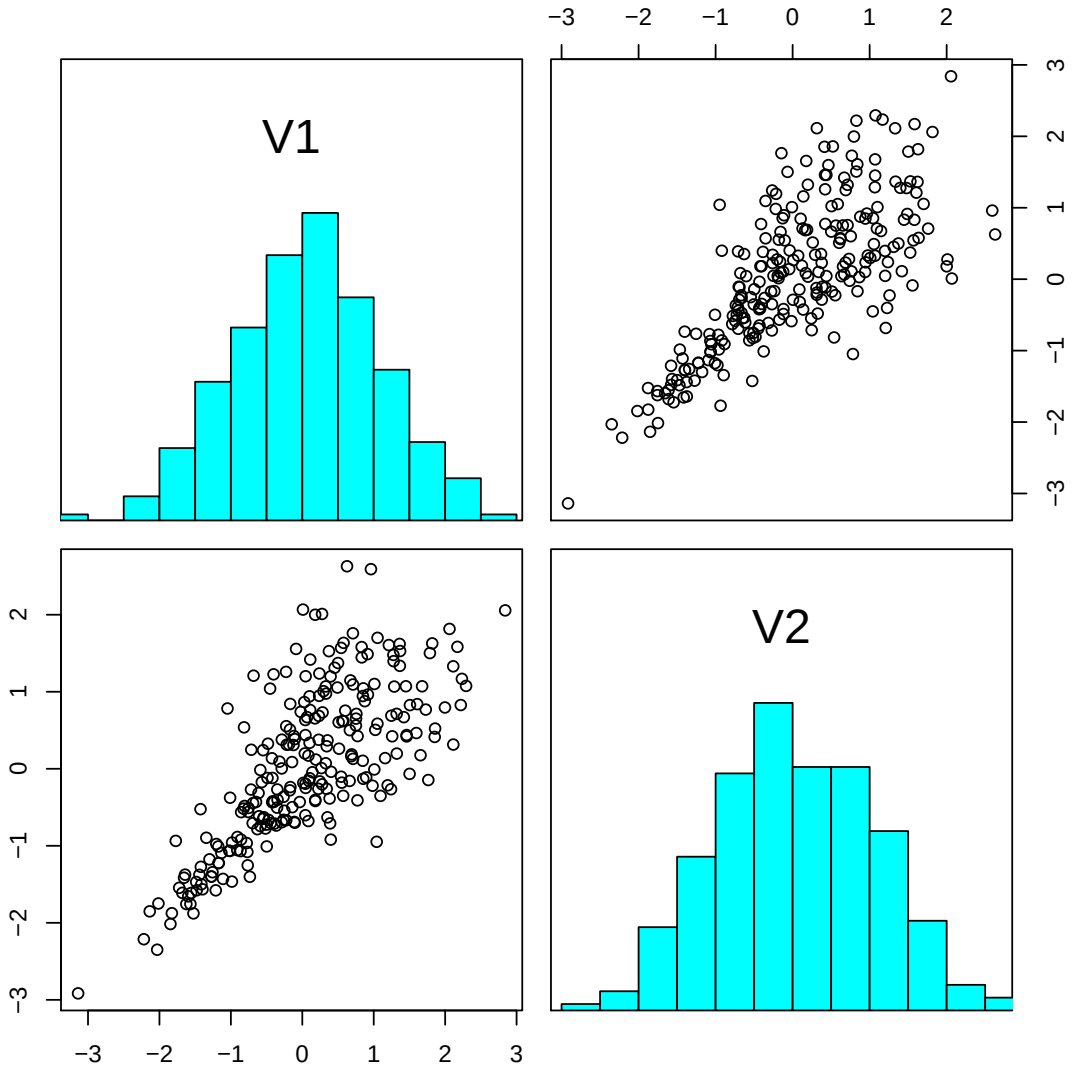

Figure 4: Bivariate scatter plot and histograms of the margins.

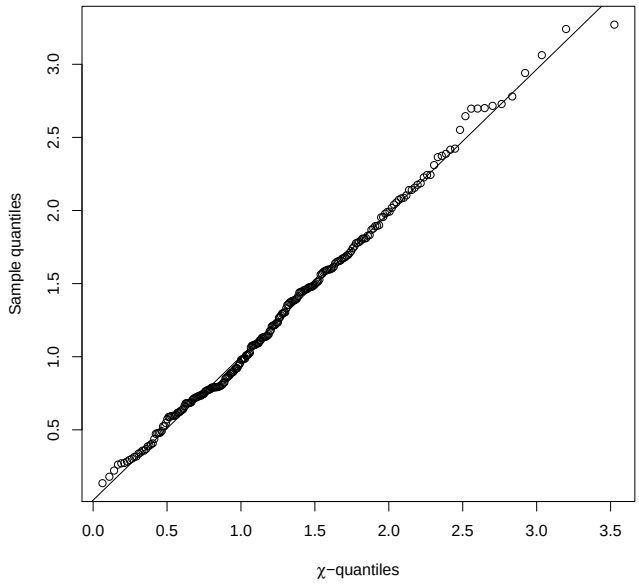

Figure 5: Q-Q plot of the Mahalanobis distances.

3. **a)** The test statistic is given by $T^2 = n \cdot d^2(\underline{\overline{x}}, \underline{0}; \mathbf{S})$. Under $H_0 : \underline{\mu} = 0$, the statistic $\frac{n-m}{m(n-1)}T^2$ has an F-distribution with m and $n-m$ degrees of freedom. Here, $m = 2$ and $n = 250$.

b) See full R code. The value is 0.7017.

c) The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true, i.e., in R, it is obtained via `1 - pf(q=(n-m)/m/(n-1)*T2,df1=m,df2=n-m)`. The p-value is 0.7054, which is larger than 0.05, hence we cannot reject H_0 on the 5% level.