Homework 2

Covariance and correlation coefficient

In statistics and probability theory, the covariance is a measure of how two random variables vary jointly from their expected values (means). When two random variables X and Y are observed over a sample of size n with measurements $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ respectively, their covariance is calculated using the formula

$$\mathrm{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

where $\bar{X}$ and $\bar{Y}$ are the respective sample means of the variables. The covariance indicates the direction of the linear relationship between the variables.

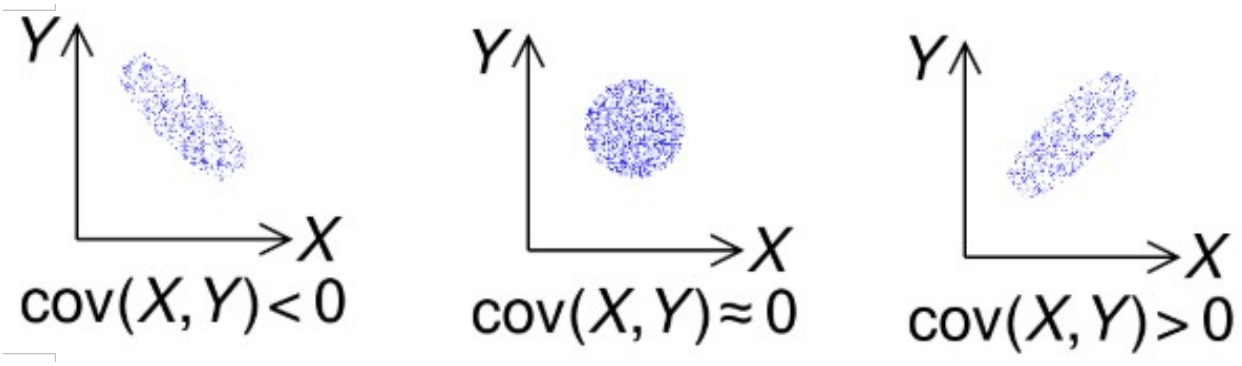

Figure 1: Covariance as as indicator of the direction of the relationship between two numeric variables.

Since the covariance depends on the scales of the variables measured, it's only useful for determining the direction of the relationship between them. To be able to also quantify the strength of a linear relationship between the variables, the correlation coefficient can be calculated using the formula

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{s_X s_Y}$$

where s_X and s_Y are the sample standard deviations, given by

$$s_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}, \quad s_Y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2}.$$

The correlation coefficient $\rho_{X,Y}$ takes values in the interval $[-1, 1]$ with values close to -1 indicating a strong negative linear relationship and values close to 1 indicating a strong positive linear relationship. Notice that the correlation coefficient being close to -1 or 1 is only a necessary, but not a sufficient condition to conclude that the relationship between the variables is linear, nor it can detect the strength non-linear relationships. For this reason the correlation analysis should be supported by some type of graphical representation of the data.

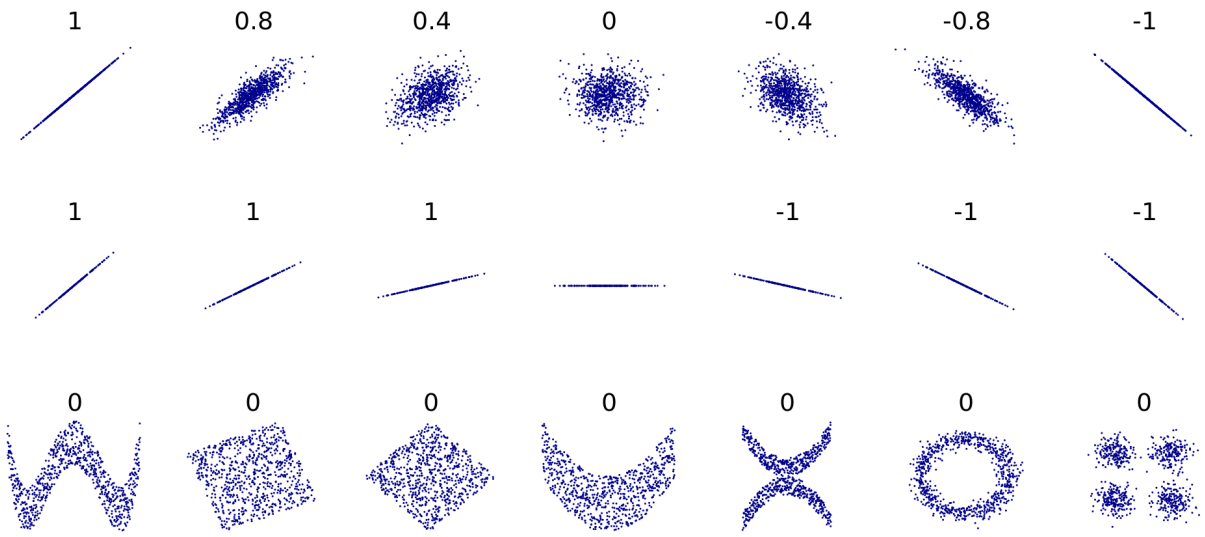

Figure 2: Several sets of (X, Y) points, with the correlation coefficient of X and Y for each set.

1. Implement the function `sample_covariance()` which returns the sample covariance and takes the following arguments:

   - `x` - a numerical vector, containing the measurements of a variable over a sample
   - `y` - a numerical vector, containing the measurement of another variable over a sample
   - `na.rm` - a boolean, for flagging whether potential `NA` values in `x` and `y` should be omitted from the calculation with a default value of `FALSE`. Note that if at least one of X_k or Y_k is `NA` for any `k`, this means that the entire k-th pair should be omitted from the covariance calculation.

   Use appropriate flow control structures to check if the arguments are of the required type and that `x` and `y` are of equal length. You may use the `mean()` function in your calculations.

2. Implement the function `sample_correlation()` which returns the sample correlation coefficient and takes the same arguments `x`, `y` and `na.rm` . Use the `sample_covariance()` function from 1. to calculate the covariance from the correlation coefficient formula.

3. Load the `diamonds` data set from the `ggplot2` package and read the description of its contents. Use the `summary()` function to display descriptive statistics for the columns.

4. Using `dplyr`'s syntax, filter the data set so that only diamonds with `price` greater than or equal to 600 remain.

5. Working with the filtered data from now on, use `dplyr` to summarize the data and calculate the average `price` per `cut`. Which `cut` level has the highest and which the lowest average `price`?

6. Using `dplyr`'s syntax create two new variables `log_carat` which is equal to ln(`carat`) and `log_price` which is equal to ln(`price`). Calculate the correlation coefficient of `log_carat` and `log_price` using the `sample_correlation()` function implemented in 2. Would you say there is a strong linear relationship between the variables?

7. Use `ggplot2` to make a scatter plot of `log_carat` against `log_price` only for the diamonds of `clarity` levels `I1` and `IF` where the different points are colored by the clarity levels. Also include smooth lines going through the data points of the two clarity levels. Would you say that the diamonds from the two groups differ?