# MATH1712 Probability and Statistics II
# Homework 3

http://www1.maths.leeds.ac.uk/~voss/2017/MATH1712/
Jochen Voss, J.Voss@leeds.ac.uk
2017/18, semester 2

*This exercise sheet will be discussed in the tutorials of the week beginning 12th February.*

The main topics of this exercise sheet are covariances and correlations. Last semester you learned about covariance and correlation for random variables: the definitions are

$$\mathrm{Cov}(X,Y) = \mathbb{E}\Big(\big(X-\mathbb{E}(X)\big)\big(Y-\mathbb{E}(Y)\big)\Big) \quad \text{and} \quad \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

This semester we will learn about sample covariance and sample correlation. For paired samples $(x_1,y_1),\dots,(x_n,y_n)$, the sample covariance and correlation are

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \quad \text{and} \quad r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}},$$

respectively, where $\bar{x}$ and $\bar{y}$ are the sample averages and $s_x^2$ and $s_y^2$ are sample variances.

**During the tutorial.** If you have a laptop which you can run R on, please bring this laptop with you to the tutorial class.

**(1)** For $n \in \{1,2,\dots,9\}$, consider the datasets from

`http://www1.maths.leeds.ac.uk/~voss/2017/MATH1712/ex03-`$n$`.csv`

Each file contains two columns, $x$ and $y$. For each dataset use R to create a scatter plot $x$ against $y$ (see overleaf) and compute the sample correlation between $x$ and $y$. Discuss how well you could guess the sample correlation from the plots alone.

**(2)** Work together on the board to find a proof that the covariance can be written as

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Discuss how to best write down this proof.

**(3)** Let $x_i = i$ for $i \in \{1,\dots,100\}$. Using R, compute the sample variance of $x_1,\dots,x_{100}$. Can you obtain the same result analytically?

**Homework questions.** Your solutions to these questions contribute towards your final mark for the module. Please hand in your solutions **to your tutor** via the silver pigeon holes (down the stairs from the maths reception) by **Monday, 19th February, 4pm**.

**Exercise 9.** Show that the sample covariance $s_{xy}$ can be written as

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \frac{n}{n-1}\bar{x}\bar{y}.$$

**Exercise 10.** The *mean squared error* of an estimator $\hat{\theta}(x_1,\dots,x_n)$ is defined to be

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\Big(\big(\hat{\theta}(X_1,\dots,X_n)-\theta\big)^2\Big),$$

where $X_1,\dots,X_n$ are a random sample from the model, using the parameter value $\theta$.

a) In one or two sentences, explain what it means for $\hat{\theta}$ when $\mathrm{MSE}(\hat{\theta})$ is small.

b) Show that $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}\big(\hat{\theta}(X_1,\dots,X_n)\big) + \big(\mathrm{bias}(\hat{\theta})\big)^2$.

c) Let $m(x_1,\dots,x_n)$ denote the sample average of $x_1,\dots,x_n$. We have seen that $m$ can be used as an estimator for the mean $\mu$. Determine $\mathrm{MSE}(m)$.

**Exercise 11.** In real problems we often have data with units, *e.g.* we could come into a situation where we would need to compute the correlation between $x = (0.47\,\text{m}, 1.38\,\text{m}, 1.39\,\text{m}, 1.5\,\text{m}, 2.1\,\text{m})$ and $y = (0.02\,\text{m}, 1.02\,\text{m}, 0.53\,\text{m}, 0.43\,\text{m}, 1.4\,\text{m})$. This is done by just stripping off the units and computing the sample correlation between the numeric values.

a) Compute the sample correlation $r_{xy}$ for the data given above.

b) Show that the sample correlation does not change when the data is converted to centimetres, *i.e.* that the rescaled vectors $\tilde{x} = (47\,\text{cm}, 138\,\text{cm}, 139\,\text{cm}, 150\,\text{cm}, 210\,\text{cm})$ and $\tilde{y} = (2\,\text{cm}, 102\,\text{cm}, 53\,\text{cm}, 43\,\text{cm}, 140\,\text{cm})$ have the same sample correlation $r_{\tilde{x}\tilde{y}} = r_{xy}$.

c) Give a mathematical proof that the sample correlation does not depend on the choice of units. (Part of the task here is to translate this into a mathematical statement which can be proved.)

**Exercise 12.** Let $\hat{\alpha}$ and $\hat{\beta}$ the estimated parameters in a least-squares regression problem with data $(x_1, y_1), \ldots, (x_n, y_n)$, and let $\hat{\varepsilon}_i = y_i - \hat{\alpha} - x_i\hat{\beta}$ for $i \in \{1, \ldots, n\}$ be the estimated residuals. Show that $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$.

**cor()**
If `x` and `y` are vectors of the same length, `cor(x,y)` returns the sample correlation between vectors `x` and `y`. The option `na.rm="complete.obs"` can be used to ignore any samples where either the $x$- or the $y$-value is missing. The `help(cor)` command shows more details and, in particular, lists different methods for dealing with missing values.

**plot()**
The `plot()` command produces line and scatter plots from paired, numerical data. If `x` contains a vector $(x_1, \ldots, x_n)$ and `y` contains a vector $(y_1, \ldots, y_n)$, then the command

```
plot(x, y)
```

produces a scatter plot, showing the observations $(x_i, y_i)$. There are many optional arguments, which can be used to adjust the plot. The most important ones are:

- `xlab="..."` and `ylab="..."` can be used to adjust the axis labels for the $x$- and $y$-axis, respectively.
- `xlim=c(a,b)` adjusts the plot so that the $x$-coordinate range from $a$ to $b$ is shown in the plot. This option can, for example, be used to exclude outliers from the plot.
- `pch=...` can be used to change the symbol used to represent samples in the scatter plot. Many values are possible, *e.g.* `pch=0` gives squares, `pch=2` gives triangles and `pch=5` gives diamonds.
- `asp=1` uses the same scale for for $x$ and $y$.

The command

```
plot(x, y, type="l")
```

produces a line plot, by connecting consecutive points using straight line segments. In addition to the arguments listed above, the following options can be used:

- `lwd=...` can be used to adjust the line width, *e.g.* `lwd=2` gives thicker lines, `lwd=0.5` gives thinner lines.
- `lty=1` up to `lty=6` can be used to get different forms of dashed or dotted lines.

Further details about the function `plot()` can be found using the command `help(plot)` in R. Graphics parameters used to adjust the plot are explained at `help(par)`.