EXERCISE: Testing an Hypothesis
Regarding Pearson's Data on Heights

From the evidence of Figure 1 of Lecture 1, which shows two regression lines fitted to Pearson's data, there seems to have been a significant increase in the heights of adult males from one generation to the next in late Victorian England. The object of this exercise is to test whether the increase is statistically significant.

Let x denote the height of a father and y denote the height of his son. We may assume that these heights are normally distributed with means of $E(x) = \mu_x$ and $E(y) = \mu_y$, with variances of $V(x) = \sigma_x^2$ and $V(y) = \sigma_y^2$ and with a covariance of $C(x, y) = \rho\sigma_x\sigma_x$, which allows us to write

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_x \\ \rho\sigma_x\sigma_x & \sigma_y^2 \end{bmatrix}\right). \tag{1}$$

We shall also assume that $V(x) = V(y) = \sigma^2$, and we shall endeavour to test the hypothesis that $E(x) = E(y) = \mu$, given a sample of N observations contained in the vectors $x = [x_1, \ldots, x_N]'$ and $y = [y_1, \ldots, y_N]'$. On that basis, there is

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x\iota \\ \mu_y\iota \end{bmatrix}, \sigma^2 \begin{bmatrix} I & \rho I \\ \rho I & I \end{bmatrix}\right), \tag{2}$$

where $\iota' = [1, 1, \ldots, 1]$ is the summation vector of N units.

To develop the relevant test statistics, one may begin by finding the joint distribution of the means of the two variables, which are given by

$$\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} = \frac{1}{N}\begin{bmatrix} \iota' & 0 \\ 0 & \iota' \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}. \tag{3}$$

You are asked to demonstrate that

$$\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \frac{\sigma^2}{N}\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right). \tag{4}$$

From this, you may derive a random variable, based on the difference between $\bar{y}$ and $\bar{x}$, that has a standard normal distribution and another equivalent variable that has a chi-square distribution.

To implement a test of the hypothesis that $\mu_x = \mu_y$, it is necessary to provide values for ρ and for σ^2. The following estimates are appropriate:

$$\hat{\sigma}^2 = \frac{1}{2}\left\{\sum_{i=1}^{N}\frac{(x_i - \bar{x})^2}{N-1} + \sum_{i=1}^{N}\frac{(y_i - \bar{y})^2}{N-1}\right\} = \frac{1}{2}\left\{\hat{\sigma}_x^2 + \hat{\sigma}_y^2\right\}, \tag{5}$$

$$\hat{\rho} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}. \tag{6}$$

Using these, a test statistic can be devised that is distributed asymptotically as a standard normal variate and a related one that is asymptotically distributed as a chi-square variate. To test the hypothesis, it should be straightforward to perform the necessary calculations by taking the data that is provided on the web site and by using an Excel spreadsheet.

Some justification may be sought for using $\bar{x}$, $\bar{y}$ and $\hat{\sigma}^2$ defined in (5) as the means of estimating μ_x, μ_q and σ^2, respectively. To this end, you should derive the maximum-likelihood estimators of these parameters for comparison.