EXERCISE 1
Suppose \( Y_i = \mu + \epsilon_i \), with \( E(\epsilon_i) = 0 \), \( \text{var}(Y_i) = \sigma^2 \), and \( \text{cov}(Y_i, Y_j) = \sigma^2 \rho^{|i-j|} \). The inverse of the variance covariance matrix is given as follows:

\[
\Sigma^{-1} = \frac{1}{\sigma^2(1 - \rho^2)} \begin{pmatrix}
1 & -\rho & 0 & 0 & \ldots & 0 \\
-\rho & 1 + \rho^2 & -\rho & 0 & \ldots & 0 \\
0 & -\rho & 1 + \rho^2 & -\rho & \ldots & 0 \\
0 & 0 & -\rho & 1 + \rho^2 & \ldots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots \\
0 & 0 & \ldots & 0 & -\rho & 1 \\
\end{pmatrix}
\]

a. It is given that \( \hat{\mu} = \frac{1}{\sum Y_i} \). Show that \( \text{var}(\hat{\mu}) = \sigma^2 \frac{1 + \rho}{(n-2)(1-\rho^2)\rho} \). Explain what happens to \( \text{var}(\hat{\mu}) \) when \( \rho = 0 \) and when \( \rho = 1 \).

b. For the same sample size \( n \), the variance of \( \hat{\mu} \) for this model (correlated data) is less precise than \( \text{var}(\bar{Y}) = \frac{\sigma^2}{n} \) in the i.i.d. case. Show that \( \text{var}(\hat{\mu}) \) is more variable than \( \text{var}(\bar{Y}) \) by a factor of \( \left[ \frac{1-\rho}{1+\rho} + \frac{2}{n} \frac{\rho}{1+\rho} \right]^{-1} \).

EXERCISE 2
Use R to access the Maas river data. These data contain the concentration of lead and zinc in ppm at 155 locations at the banks of the Maas river in the Netherlands. You can read the data in R as follows:

```r
a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics_c173_c273/soil.txt", header=TRUE)
```

a. Compute the summary statistics for lead and zinc.

b. Plot the histogram of lead and log(lead).

c. Plot log(lead) against log(zinc). What do you observe?

d. The level of risk for surface soil based on lead concentration in ppm is given on the table below:

<table>
<thead>
<tr>
<th>Mean concentration (ppm)</th>
<th>Level of risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below 150</td>
<td>Lead-free</td>
</tr>
<tr>
<td>Between 150-400</td>
<td>Lead-safe</td>
</tr>
<tr>
<td>Above 400</td>
<td>Significant environmental lead hazard</td>
</tr>
</tbody>
</table>

Use similar R commands as in pages 8-11 of handout #4 to create a bubble plot with different colors and sizes of the lead concentration at these 155 locations.
Exercise 3
Consider the equal correlation model, $\text{cov}(Y_i, Y_j) = \rho \sigma^2$. The variance covariance matrix is of the form $(a - b)I + bJ$, where $a = 1, b = \rho, J = 11'$. Therefore, in our model $\Sigma = \sigma^2 [(1 - \rho)I + \rho J]$. The inverse of this special matrix can be obtained as follows: $\Sigma^{-1} = \frac{1}{\sigma^2 (1 - \rho)} \left[ I - \frac{\rho}{1 + [(n - 1) \rho]} J \right]$. The GLS estimator of $\mu$ is $\hat{\mu} = \frac{1}{\Sigma} \Sigma^{-1} Y$ with variance $\text{var}(\hat{\mu}) = \frac{1}{\Sigma} \Sigma^{-1}$ and using the inverse of the variance covariance matrix above we get $\text{var}(\hat{\mu}) = \frac{\sigma^2}{n} \left[ 1 + \frac{1}{n} \frac{n}{n - 1} \right]$. Show that the variance of the error of prediction is $\sigma^2_{\text{pred}} = \sigma^2 \left[ 1 + \frac{1}{n} \frac{n}{n - 1} \right]$. Find the constant $c$ (it will be a function of $n$ and $\rho$) and find the condition on $\rho$ that will make the prediction variance $\sigma^2_{\text{pred}}$ in the correlated data more precise than the $\text{var}(Y_0 - \bar{Y}) = \sigma^2 (1 + \frac{1}{n})$ in the i.i.d. case.

Exercise 4
Suppose $Y_1, Y_n$ are normally distributed with $E(Y_i) = \mu, \text{var}(Y_i) = \sigma^2, \text{cov}(Y_i, Y_j) = \rho \sigma^2$. Define $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. Are $\bar{Y}, S^2$ independent? Is $S^2$ an unbiased estimator of $\sigma^2$?

Exercise 5
Assume that $Y_1, \ldots, Y_n$ follow the normal distribution with mean $\mu_Y$, variance $\sigma^2$, and $\text{cov}(Y_i, Y_j) = \rho \sigma^2$. Consider a second sample such that $X_1, \ldots, X_n$ follow the normal distribution with mean $\mu_X$, variance $\sigma^2$, and $\text{cov}(X_i, X_j) = \rho \sigma^2$. The two samples are independent.

a. Show that $\text{var}(\bar{Y}) = \frac{\sigma^2}{n} [1 + (n - 1) \rho]$.

b. Assume that $\rho > 0$. Which sample mean is more dispersed? The one that takes into account the correlation or the one that ignores correlation (i.i.d. case).

c. Consider the case with the correlated $Y_i$’s. Find $E\bar{Y}$. Is $\bar{Y}$ unbiased estimator of $\mu_Y$? Find $\lim_{n \to \infty} \text{var}(\bar{Y})$. Is $\bar{Y}$ consistent estimator of $\mu$?

d. Assume that $\sigma^2$ is known and that the data are uncorrelated. Write the test statistic for testing the hypothesis $H_0 : \mu_x = \mu_y$.

e. Refer to question (d). Assume now that the data are correlated. Write the test statistic for testing the hypothesis $H_0 : \mu_x = \mu_y$.

f. Compare (d) with (e). Which test has smaller $p$-value? Which test rejects more often than it should?

g. If $n$ denotes the sample size of the correlated observations find the corresponding sample size of uncorrelated observations.

h. Consider a confidence interval for $\mu_x$ using the correlated data of the second sample: $\bar{x} \pm 1.96 \sigma \sqrt{\frac{1 + (n - 1) \rho}{n}}$. Now, suppose we fail to see the dependence in the random variables, and we decided to use $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$. What is the actual coverage of our confidence interval if $n = 25, \rho = 0.2$?