University of California, Los Angeles
Department of Statistics

Statistics C173/C273 **Instructor: Nicolas Christou**

Homework 1

EXERCISE 1
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:

```
soil <- 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:

Mean concentration (`ppm`)	Level of risk
Below 150	Lead-free
Between 150-400	Lead-safe
Above 400	Significant environmental lead hazard

Use techniques similar to pages 8-11 of the first handout to give different colors and sizes to the lead concentration at these 155 locations.

EXERCISE 2
The data for this exercise represent approximately the centers (given by longitude and latitude) of each one of the City of Los Angeles neighborhoods. See also the Los Angeles Times project on the City of Los Angeles neighborhoods at:

`http://projects.latimes.com/mapping-la/neighborhoods/`

You can access these data at:

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

a. Plot these data points and add the map on the plot.

b. Do you see any relationship between income and school performance? Hint: Plot the variable `Schools` against the variable `Income` and describe what you see. Also, ignore the data points on the plot for which `Schools=0`.

EXERCISE 3
The Wolfcamp aquifer data (see Cressie 1993, pp. 212-214). In the 1980s the U.S. Department of Energy proposed a nuclear waste site to be in Deaf Smith County in Texas (bordering New Mexico). The contamination of the aquifer was a concern, and therefore the piezometric-head data were obtained at 85 locations by drilling a narrow pipe through the aquifer. The measures are in feet above sea level. You can access the data here:

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

a. Use `R` to run the regression of `level` on the coordinates `x` and `y`. What do the negative signs of the estimated coefficients mean?

b. Convert the data frame `a` into a geodata object using `geoR`. Use the `points` function to support your answer to (a).

EXERCISE 4

The sample mean $\bar{X}$ is a good estimator of the population mean μ. It can also be used to predict a future value of X independently selected from the population. Assume that you have a sample mean $\bar{x}$ and a sample variance s^2, based on a random sample of n measurements from a normal population $(X_1, X_2, \ldots, X_n)$. Construct a prediction interval for a new observation x, say x_p. Use $1 - \alpha$ confidence level. Hints:

a. Find the distribution of the random quantity $X_p - \bar{X}$.

b. It is known that $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$.

c. Use the definition of the t distribution to obtain the prediction interval: $t = \frac{Z}{\sqrt{\frac{\chi^2_{df}}{df}}}$, where $Z \sim N(0, 1)$.

EXERCISE 5

Consider the following simple regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for which $E(\epsilon_i) = 0$, $E(\epsilon_i \epsilon_j) = 0$ for $i \neq j$, and $\text{var}(\epsilon_i) = \sigma^2$. Show that $Cov(\bar{Y}, \hat{\beta}_1) = 0$ where $\bar{Y}$ is the sample mean of the y values, and $\hat{\beta}_1$ is the estimate of β_1.

EXERCISE 6

Let $X_1, X_2, \ldots, X_n$ be a random sample from normal distribution with $E(X_i) = \mu, var(X_i) = \sigma^2$, and $cov(X_i, X_j) = \rho\sigma^2$, for $i \neq j$. Then $\bar{X} \sim N(\mu, \sigma\sqrt{\frac{1+(n-1)\rho}{n}})$, and a 95% confidence interval for μ will be

$$\bar{x} \pm 1.96\sigma\sqrt{\frac{1 + (n-1)\rho}{n}}. \tag{1}$$

Note: If $\rho = 0$ then we get the usual confidence interval for μ when we are dealing with i.i.d. random variables i.e.,

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}. \tag{2}$$

Now, suppose we fail to see the dependence in our random variables, and instead of using (1) we decided to use (2). What is the actual coverage of our confidence interval if $n = 25, \sigma = 3, \rho = 0.2$, and $\bar{x} = 20$?