University of California, Los Angeles
Department of Statistics

Statistics 13 **Instructor: Nicolas Christou**

Lab 1

Note: All the graphs of this lab should be saved as pdf files and be inserted in a Word document where you can type your comments. Make sure you write a comment for each graph!

EXERCISE 1

The data here represent life expectancies (`Life`) and per capita income (`Income`) in 1974 dollars for 101 countries in the early 1970's. The source of these data is: Leinhardt and Wasserman (1979), *New York Times* (September, 28, 1975, p. E-3). They also appear on *Regression Analysis* by Ashish Sen and Muni Srivastava. You can access these data in `R` using:

```
c <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/countries_life.txt", header=TRUE)
```

a. Construct a scatterplot of `Life` against `Income`. Note: `Income` should be on the horizontal axis. How does income affect life expectancy?
b. Construct the boxplot and histogram of `Income`. Are there any outliers?
c. Split the data set into two part: One for which the `Income` is below $1000, and one for which the `Income` is above $1000.
d. Use the data for which the `Income` is below $1000: Plot `Life` against `Income` and compute the correlation coefficient.

EXERCISE 2

These data represent the percentage of body fat determined by underwater weighing and various body circumference measurements and other variables for 251 men. For the variable description see the handout "Data analysis with R":

```
http://www.stat.ucla.edu/~nchristo/statistics13/stat13_intro_to_R.pdf
```

You can access the data in `R` with the following command:

```
d <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/body_fat.txt", header=TRUE)
```

a. Compute summary statistics for the variable y (percent body fat).
b. Construct the boxplot and histogram of y.
c. Create two data sets: one for which $x_3 < 30$ (men younger than 30) and one for which $x_3 > 60$ (men older than 60). Construct side-by-side boxplots of y for these two data sets.
d. Compute summary statistics for y for the two data sets that you created in part (c). What do you observe?
e. Compute the correlation coefficient between y and x_{10} for the two data sets that you created in part(c).

EXERCISE 3

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:

```
a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/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 9, 10, and 11 in the handout "Introduction to `R`" to give different colors and sizes to the lead concentration at these 155 locations.

EXERCISE 4

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:

```
b <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/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`.