University of California, Los Angeles
Department of Statistics

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

Homework 2

EXERCISE 1

Suppose $Y_i = \mu + \epsilon_i$, with $E(\epsilon_i) = 0, \text{var}((Y_i) = \sigma^2$, and $\text{cov}(Y_i, Y_j) = \sigma^2 \rho^{|i-j|}$. Assume that $\rho > 0$. Draw the power function in R for testing $H_0 : \mu = \mu_0$ against $H_a : \mu > \mu_o$ when we use the test statistic $Z = \frac{\bar{Y}-\mu_0}{\text{sd}(\bar{Y})}$. On the same plot, draw also the power function when $\rho = 0$ when testing the same hypothesis.

EXERCISE 2

In homework 1, exercise 1(b) we showed that the relative excess variability (REV) for $\hat{\mu}$ in the non-i.i.d. case is given by $\left[\frac{1-\rho}{1+\rho} + \frac{2}{n}\frac{\rho}{1+\rho}\right]^{-1}$. For the same sample size n, the variance of $\hat{\mu}$ in the correlated data is less precise than the variance of $\hat{\mu} = \bar{Y}$ in the i.i.d. case. Use `R` to construct a table that shows this extra variability in $var(\hat{\mu})$ by using
$\rho = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$, and $n = 1, 2, 3, \ldots, 50$.

EXERCISE 3

Suppose $Y_1, \ldots, Y_n$ are i.i.d. random variables with $Y_i \sim N(\mu, \sigma)$. Predict Y_0 using the method of Lagrange multiplier. Note: Assume that $\hat{Y}_0 = \sum_{i=1}^n w_i y_i$ with $E\hat{Y}_0 = \mu$ and minimize $\text{var}(Y_0 - \hat{Y}_0)$. The answer should be $w_i = \frac{1}{n}, i = 1, \ldots, n$ and therefore $\hat{Y}_0 = \bar{Y}$.

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:

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

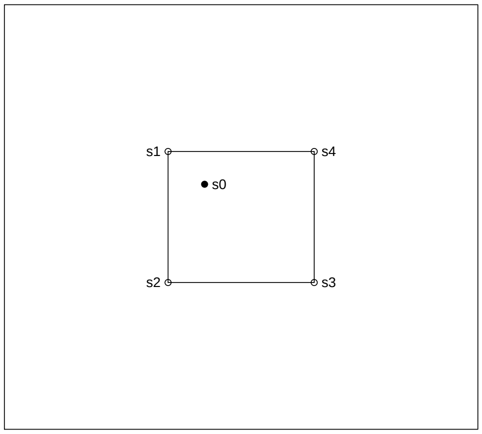
a. Plot these data points and add the map on the plot.

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. (Ignore the data points on the plot for which `Schools=0`).

EXERCISE 5

From week 2 - lecture 2: We assume that $Y_1, \ldots, Y_n$ follows a multivariate normal distribution, i.e. $\mathbf{Y} \sim N(\mu\mathbf{1}, \sigma^2\mathbf{V})$. The maximum likelihood estimate of σ^2 is given by $\hat{\sigma}^2 = \frac{(\mathbf{Y}-\hat{\mu}\mathbf{1})'\mathbf{V}^{-1}(\mathbf{Y}-\hat{\mu}\mathbf{1})}{n}$, where $\hat{\mu} = \frac{\mathbf{1}'\mathbf{V}^{-1}\mathbf{Y}}{\mathbf{1}'\mathbf{V}^{-1}\mathbf{1}}$. Use properties of the trace to find $E[\hat{\sigma}^2]$.

EXERCISE 6
Let $Z(s)$ be a stationary random function, and let $\hat{Z}(s_0) = 0.5Z(s_1), +0.2Z(s_2)+0.2Z(s_3)+0.1Z(s_4)$ be a weighted average of the four values $Z(s_1), Z(s_2), Z(s_3), Z(s_4)$ as shown on the $2m \times 2m$ square below.



Answer the following questions:

a. Use `R` to compute the distance matrix.

b. Assume that the spatial covariance is given by $C(h) = 2.5e^{-\frac{h}{2}}$. Use `R` to compute the 4×4 variance covariance matrix of the vector $\mathbf{Z} = \begin{pmatrix} Z(s_1) \\ Z(s_2) \\ Z(s_3) \\ Z(s_4) \end{pmatrix}$.

c. Use matrix and vector notation to compute $var(\hat{Z}(s_0))$. Note: $\hat{Z}(s_0) = \mathbf{a}'\mathbf{Z}$, where $\mathbf{a} = \begin{pmatrix} 0.5 \\ 0.2 \\ 0.2 \\ 0.1 \end{pmatrix}$. Then use $var(\hat{Z}(s_0)) = \mathbf{a}'\mathbf{\Sigma}\mathbf{a}$.

EXERCISE 7
Access the Kruger Park data:

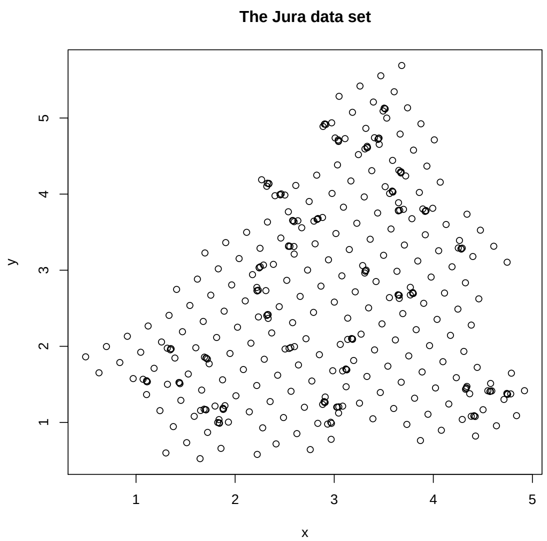
```
a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics_c173_c273/
kruger_park_rainfall.txt", header=TRUE)
```

Load the packages `sp` and `gstat` to construct h-scatterplots with separation distances from 0 to 1 by 0.1.

EXERCISE 8
Access the following data:

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

These Jura data were collected by the Swiss Federal Institute of Technology at Lausanne. See Goovaerts, P. 1997, "Geostatistics for Natural Resources Evaluation", Oxford University Press, New-York, 483 p. for more details. Data were recorded at 359 locations scattered in space (see figure below).



Concentrations of seven heavy metals (cadmium, cobalt, chromium, copper, nickel, lead, and zinc) in the topsoil were measured at each location. The type of land use and rock type was also recorded for each location.

```
> names(a)
 [1] "x"        "y"        "Landuse" "Rock"     "Cd"
 [6] "Co"       "Cr"       "Cu"      "Ni"       "Pb"
[11] "Zn"

y <- a$Pb

x1 <- a$Cd
x2 <- a$Co
x3 <- a$Cr
x4 <- a$Cu
x5 <- a$Ni
x6 <- a$Zn
```

The variables x, y are the coordinates. Landuse and Rock represent type of land use (forest, pasture, meadow, tillage) and rock type (Argovian, Kimmeridgian, Sequanina, Portlandian, and Quaternary). The other variables are concetrations in ppm of the following chemical elements: `Cd`: Cadmium, `Co`: Cobalt, `Cr`: Chromium, `Cu`: Copper, `Ni`: Nickel, `Pb`: Lead, `Zn`: Zinc.

Construct a circle plot for one of these chemical elements.