STAT 636, Fall 2017 - Assignment 1
Due Tuesday, September 12, 11:55pm Central
Submit your assignment through eCampus.

1. The `Oxygen` data contain $p = 4$ oxygen volume measurements for 25 males and 25 females. The variables are $X_1$: oxygen volume (L/min.) while resting, $X_2$: oxygen volume (mL/kg/min.) while resting, $X_3$: oxygen volume (L/min.) during strenuous exercise, and $X_4$: oxygen volume (mL/kg/min.) during strenuous exercise. You can use the following `R` code to load these data:

```
dta <- read.delim("oxygen.DAT", header = FALSE, sep = "")
colnames(dta) <- c("X_1", "X_2", "X_3", "X_4", "Gender")
```

   (a) Report a table showing the sample averages and standard deviations for each variable, by gender. Comment.

   (b) Make a pairs plot like we did for the pottery data. Comment on any relationships you see. Which individual would you say is an outlier?

   (c) Make a coplot, like we did for the pottery data, to compare $X_1$ to $X_3$ by gender. Does there appear to be a difference for this pair of variables between genders?

2. The multivariate normal distribution is defined by its probability density function (pdf)

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}$$

for $-\infty < x_i < \infty$, $i = 1, 2, \ldots, p$. Volumes underneath this surface equal probabilities. The mean *vector* of this distribution is $\boldsymbol{\mu}$, and the covariance *matrix* is $\mathbf{\Sigma}$. In the bivariate setting ($p = 2$), we have

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

Thus, draws from this distribution are *pairs* (vectors of length $p = 2$). The averages of the two components among all pairs in the population equal $\mu_1$ and $\mu_2$, respectively. Similarly, the variances of the two components among all pairs in the population equal $\sigma_{11}$ and $\sigma_{22}$, respectively. Finally, the *covariance* between the two components equals $\sigma_{12}$, which means that the correlation equals $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$.

For the bivariate normal distribution:

   (a) Recall that the distance between the point $P = (x_1, x_2)$ and $Q = (\mu_1, \mu_2)$ can be written as

$$d(P, Q) = \sqrt{a_{11}(x_1 - \mu_1)^2 + 2a_{12}(x_1 - \mu_1)(x_2 - \mu_2) + a_{22}(x_2 - \mu_2)^2}$$

   We will see that the statistical distance between the two *vectors* $\mathbf{x}$ and $\boldsymbol{\mu}$ can be written as

$$d(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

   And it turns out that $d(P, Q) = d(\mathbf{x}, \boldsymbol{\mu})$. Use this result to derive the values of $a_{11}$, $a_{12}$, and $a_{22}$.

(b) Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & -1.6 \\ -1.6 & 4.0 \end{pmatrix}$$

i. Use the `persp` function to graph the pdf. You can do this by evaluating the pdf over a grid of $\mathbf{x}$ values. Code the pdf manually; i.e., do not use the `dmvnorm` function (or any other predefined function). Use the `ticktype = ‘‘detailed’’` option to include axis tick marks and labels.
Here are some `R` functions and operations that you will find useful: `sqrt` computes the square root; `det` computes the determinant of a matrix; `t(v)` computes the transpose of the vector / matrix `v`; `u %*% v` computes the vector / matrix product of the vectors / matrices `u` and `v`; `exp(a)` equals $e^a$; `solve` inverts a matrix. If you need a refresher on the vector / matrix operations, see the textbook.

ii. Let $O$ be the origin $(0, 0)$, $P$ be the point $(0, -2)$, and $Q$ be the point $(\mu_1, \mu_2) = (1, -1)$. Which of $O$ or $P$ is "closer" to $\boldsymbol{\mu}$, based on statistical distance? Which of $O$ or $P$ is closer to $\boldsymbol{\mu}$, based on straight-line distance?

iii. Consider all of the pairs $(x_1, x_2)$ located inside a small square centered at $O$. That is, let $R_O$ be the square containing all pairs $(x_1, x_2)$ such that $-\epsilon \le x_1 \le \epsilon$ and $-\epsilon \le x_2 \le \epsilon$ for some small value of $\epsilon$ (e.g., $\epsilon = 0.01$). Similarly, let $R_P$ consist of all pairs located inside an equally-small square centered at $P$, for which $-\epsilon \le x_1 \le \epsilon$ and $-2 - \epsilon \le x_2 \le -2 + \epsilon$. Let $P(\mathbf{x} \in R_O)$ be the probability that a randomly-drawn pair from this bivariate normal distribution falls within $R_O$. Similarly, let $P(\mathbf{x} \in R_P)$ be the probability that a randomly-drawn pair falls within $R_P$. Is $P(\mathbf{x} \in R_O) < P(\mathbf{x} \in R_P)$, $P(\mathbf{x} \in R_O) = P(\mathbf{x} \in R_P)$, or $P(\mathbf{x} \in R_O) > P(\mathbf{x} \in R_P)$? Why? No calculations are required to answer this.

3. Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & -2 \end{bmatrix}$$

Without using a computer:

(a) Find the eigenvalues and normalized eigenvectors of $\mathbf{A}$.

(b) Write the spectral decomposition of $\mathbf{A}$.

(c) Verify that the determinant of $\mathbf{A}$ equals the product of its eigenvalues.

(d) The trace of a square matrix equals the sum of its diagonal elements. Verify that the trace of $\mathbf{A}$ equals the sum of its eigenvalues.

(e) Is $\mathbf{A}$ orthogonal? Why or why not?

(f) Is $\mathbf{A}$ positive definite? Why or why not?

(g) Find $\mathbf{A}^{-1}$ and determine its eigenvalues and normalized eigenvectors.

4. Consider the matrices

$$\mathbf{A} = \begin{bmatrix} 4.000 & 4.001 \\ 4.001 & 4.002 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 4.000 & 4.001 \\ 4.001 & 4.002001 \end{bmatrix}$$

These matrices are identical except for a small difference in the (2, 2) position. Also, the columns of $\mathbf{A}$ and $\mathbf{B}$ are nearly linearly dependent. Show that $\mathbf{A}^{-1} \approx (-3)\mathbf{B}^{-1}$. So, small changes - perhaps due to rounding - can result in substantially different inverses.

5. Derive expressions for the means and variances of the following linear combinations in terms of the means and covariances of the random variables $X_1$, $X_2$, and $X_3$.

   (a) $X_1 - 2X_2$.

   (b) $X_1 + 2X_2 - X_3$.

   (c) $3X_1 - 4X_2$ if $X_1$ and $X_2$ are independent (so, $\sigma_{12} = 0$).

6. Let $\boldsymbol{\mu}' = [1, 1]$, and consider the following covariance matrices

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.00 & 0.80 \\ 0.80 & 1.00 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix} \quad \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1.00 & -0.80 \\ -0.80 & 1.00 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 1.00 & 0.40 \\ 0.40 & 0.25 \end{bmatrix} \quad \boldsymbol{\Sigma}_5 = \begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 0.25 \end{bmatrix} \quad \boldsymbol{\Sigma}_6 = \begin{bmatrix} 1.00 & -0.40 \\ -0.40 & 0.25 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_7 = \begin{bmatrix} 0.25 & 0.40 \\ 0.40 & 1.00 \end{bmatrix} \quad \boldsymbol{\Sigma}_8 = \begin{bmatrix} 0.25 & 0.00 \\ 0.00 & 1.00 \end{bmatrix} \quad \boldsymbol{\Sigma}_9 = \begin{bmatrix} 0.25 & -0.40 \\ -0.40 & 1.00 \end{bmatrix}$$

   For each covariance matrix:

   (a) Draw the ellipse consisting of all points $\mathbf{x}' = [x_1, x_2]$ for which

$$(\mathbf{x} - \boldsymbol{\mu})' \, \boldsymbol{\Sigma}^{-1} \, (\mathbf{x} - \boldsymbol{\mu}) = \chi_2^2(0.05)$$

   where $\chi_2^2(0.05)$ is the 95th percentile of the chi square distribution with $p = 2$ degrees of freedom. You can draw it by hand if you want, as long as you label the axis tick marks carefully. Alternatively, you can use the `draw.ellipse` function from the `plotrix` package.

   (b) Simulate 5000 realizations from the corresponding bivariate normal distribution using `rmvnorm` function from the `mvtnorm` package and compute the proportion that are inside the ellipse you just drew.

   For an arbitrary multivariate normal random vector $\mathbf{X} = [X_1, X_2, \ldots, X_p]$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, what would you guess $P\left((\mathbf{X} - \boldsymbol{\mu})' \, \boldsymbol{\Sigma}^{-1} \, (\mathbf{X} - \boldsymbol{\mu}) \leq \chi_p^2(\alpha)\right)$ equals?

7. Consider the random vector $\mathbf{X}' = [X_1, X_2, X_3, X_4]$ with mean vector $\boldsymbol{\mu}' = [4, 3, 2, 1]$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 1 & 1 & 0 \\ 2 & 1 & 9 & -2 \\ 2 & 0 & -2 & 4 \end{bmatrix}$$

Partition $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \hdashline X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \hdashline \mathbf{X}^{(2)} \end{bmatrix}$$

Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 1 & -2 \\ 2 & -1 \end{bmatrix}$$

and consider the linear combinations $\mathbf{AX}^{(1)}$ and $\mathbf{BX}^{(2)}$. Find the following:

(a) $E\left(\mathbf{X}^{(1)}\right)$.

(b) $E\left(\mathbf{BX}^{(2)}\right)$.

(c) $\text{Cov}\left(\mathbf{AX}^{(1)}\right)$.

(d) $\text{Cov}\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right)$.

(e) $\text{Cov}\left(\mathbf{AX}^{(1)}, \mathbf{BX}^{(2)}\right)$.