STAT 636, Fall 2015 - Assignment 1
Due Monday, September 7, 11:55pm Central
Online Students: Submit your assignment through WebAssign.
On-Campus Students: Email your assignment to the TA.

1. The data in Table 6.12 of the textbook 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.

   (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 μ_1 and μ_2, respectively. Similarly, the variances of the two components among all pairs in the population equal σ_{11} and σ_{22}, respectively. Finally, the *covariance* between the two components equals σ_{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 Supplement 2A in the textbook. We will revisit them in more detail in Topic 2.

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 $-\epsilon \leq x_1 \leq \epsilon$ and $-\epsilon \leq x_2 \leq \epsilon$ for some small value of ϵ (e.g., $\epsilon = 0.01$). Similarly, let R_P consist of all pairs located inside an equally-small square centered at P, for which $-\epsilon \leq x_1 \leq \epsilon$ and $-2 - \epsilon \leq x_2 \leq -2 + \epsilon$. 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.