STAT 5572
Final
Due: Monday, Dec.11th by 11:59 pm

(1) (8 pts) 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) $X_1 + X_2 + X_3$
(b) $2X_1 - 3X_2$ if X_1 and X_2 are independent so, $\sigma_{12} = 0$.

(2) (12 pts) Consider the bivariate normal distribution. The draws from this distribution are pairs (vectors of length $p = 2$). In the bivariate setting, we have

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

In general, the statistical distance of the point $P = (x_1, x_2)$ from the fixed point $Q = (\mu_1, \mu_2)$ can be written as,

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

Statistical distance between the two vectors x and μ can be written as,

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

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

(3) (20 pts) The data file '*turnips.dat*' contains 10 observations on 3 variables: $X_1 =$ available soil calcium, $X_2 =$ exchangeable soil calcium, and $X_3 =$ calcium content in turnip greens. The variables were measured at 10 different locations.

Desirable levels for X_1 and X_2 are 15.0 and 6.0 respectively, and the expected level of X_3 is 2.85. Assuming the multivariate normality assumption is satisfied,
test the null hypothesis $H_0: \mu' = [15.0, 6.0, 2.85]$ at $\alpha = 0.05$ using the Hotelling's T^2 test. What is the test statistic, critical value, and the p-value? What is your conclusion regarding H_0?

(4) (25 pts) Peanuts are an important crop in parts of the southern United States. To develop improved plants, crop scientists routinely compare varieties with respect to several variables. The data for one two-factor experiment are given in the file '*peanuts.dat'*.
Three varieties (5, 6, and 8) were grown at two geographical locations (1, 2) and, in this case, the three variables representing yield and the two important grade-grain characteristics were measured.

The three variables are

X_1 = Yield (plot weight)

X_2 = Sound mature kernels (weight in grams-maximum of 250 grams)

X_3 = Seed size (weight, in grams of 100 seeds)

There were two replications of the experiment.

Perform a two-factor MANOVA using the data. Test for a location effect, a variety effect, and a location-variety interaction. Use $\alpha = 0.05$.

(5) The data file '*hematology.csv*' contains six hematology variables measured on 51 workers. The variables are,

X_1 = hemoglobin concentration

X_2 = packed cell volume

X_3 = white blood cell count

X_4 = lymphocyte count

X_5 = neutrophil count

X_6 = serum lead concentration

(a) (5 pts) Read the data to R using the command: read.csv("hematology.csv"). Obtain the sample var-cov matrix S.

(b) (15 pts) Conduct a principal component analysis using S or R where R is the correlation matrix. Which is more appropriate here?

(c) (5 pts) Does the large variance of X_3 affect the pattern of the components obtain using S?

(d) (5 pts) Find the percentage of variance explained by the components. How many components would you retain?

(e) (5 pts) Identify the variables which are important for the selected components. Interpret your results.