STA 4702/5701 – Spring 2017 – Exam 3 PRINT Name ____________________

Note: Conduct all tests at at α = 0.05 significance level. SHOW ALL WORK.

Q.1. A study involving e-commerce selection of sunglasses was conducted in Malaysia. There were p = 30 words used to describe n = 20 pairs of sunglasses (words like: trendy, glamorous, classic…). Subjects rated the sunglasses by the 30 words (each on a 1-5 scale). The authors were interested in describing the correlation matrix among the keywords applied to the sunglasses. Note that the correlation matrix among the words is 30x30. The 5 largest eigenvalues of the correlation matrix are given below. Give the percentage of the total variation in ratings due to each of the first 5 principal components, as well as the cumulative percentages.

	Factor1	Factor2	Factor3	Factor4	Factor5
Eigenvalue	14.51	7.14	2.37	1.09	0.82
Variability(%)					
Cumulative(%)					

Q.2. There are 2 populations of individuals: π_1 and π_2. The density functions, prior probabilities and costs of misclassification are given below.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad f_1(\mathbf{x}) = 4x_1x_2 \quad f_2(\mathbf{x}) = 4(1-x_1)(1-x_2) \quad 0 \le x_1, x_2 \le 1 \quad p_1 = 0.6 \quad C(1|2) = 2C(2|1)$$

How will individuals with the following (x_1, x_2) values be classified? A = (0.1, 0.1), B = (0.5, 0.5), C = (0.9, 0.9).

A: ________________________ B: ________________________ C: ________________________

Q.3. A multivariate multiple regression model was fit on the NFL combine data, relating Y_1 (40 Yard Time) and Y_2 (Bench Press Reps at 225 pounds) to Z_1 (Weight) and Z_2 (Height). The estimated regression coefficients and the ML estimates for the variance/covariance matrices for **Y** ($V\{\mathbf{Y}\}=\mathbf{\Sigma}$) are given for the following 2 models (n=200 players):

Model 1: $E\{Y_k\} = \beta_{0k} + \beta_{1k}Z_1 + \beta_{2k}Z_2 \quad k = 1,2$

Model 2: $E\{Y_k\} = \beta_{0k} + \beta_{1k}Z_1 \quad k = 1,2$

```
Model 1 Beta-hat     time40   bench
intercept            4.1052 33.9310
wt                   0.0062  0.1298
height              -0.0116 -0.6209

Model 1 Sigma-hat    time40   bench
time40               0.0172 -0.1565
bench               -0.1565 19.8332

Model 2 Beta-hat     time40   bench
intercept            3.3794 -5.0717
wt                   0.0057  0.1025

Model 2 Sigma-hat    time40   bench
time40               0.0177 -0.1299
bench               -0.1299 21.2662
```

p.3.a. Give the predicted 40 Yard Times and Bench Press Reps, based on Model 1 with a player that is Z_1 = 210 pounds and Z_2 = 74 inches.

40 yard Time ________________________________ Bench Press Reps ______________________________

p.3.b. Test H_0: $\beta_{21} = \beta_{22} = 0$

Test Statistic: ___________________ Rejection Region: _____________________ P-value > or < 0.05

Q.4. For the LPGA 2008 data, we define $X^{(1)}$ as the average driving distance ($X_1^{(1)}$) and fairway accuracy percent ($X_2^{(1)}$); and $X^{(2)}$ as Sand save percent ($X_1^{(2)}$) and Putts per round ($X_2^{(2)}$). These two aspects represent long and short skills. The eigenvalues of $R_{11}^{-1/2}R_{12}R_{22}^{-1}R_{21}R_{11}^{-1/2}$ are 0.06261 and 0.00192, respectively. The sample size is n = 157 golfers.

p.4.a. What is the correlation between the first canonical variates of the standardized $X^{(1)}$ and $X^{(2)}$ sets of variables? The second canonical variates of the standardized $X^{(1)}$ and $X^{(2)}$ sets of variables?

$\text{CORR}\left(\hat{U}_1, \hat{V}_1\right) =$ ______________________________ $\text{CORR}\left(\hat{U}_2, \hat{V}_2\right) =$ ______________________________

p.4.b. Test $H_0 : \boldsymbol{\Sigma}_{12} = \boldsymbol{\rho}_{12} = \mathbf{0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$

Test Statistic: ______________________ Rejection Region: ______________________ P-value $<$ or $>$ 0.05

Q.5. A principal component analysis is conducted for p = 5 variables, based on a sample correlation matrix, $\boldsymbol{R}$, based on a sample of n = 100 units. The largest eigenvalue of $\boldsymbol{R}$ is 3.2. Compute a 95% Confidence Interval for the variable for the population's largest eigenvalue of $\boldsymbol{\rho}$, λ_1.

95% Confidence Interval: ____

Q.6. A sample of n = 151 NASCAR races from the 1970s were observed and the following variables were measured for each race: X_1 = # of Drivers, X_2 = Race Length (miles), X_3 = # of Caution Flags (crashes), and X_4 = # Lead Changes. The sample correlation matrix, its eigenvalues and eigenvectors are given below.

```
> (R <- cor(race))
           drivers   racelen  cautions  leadchng
drivers  1.0000000 0.7906728 0.1599884 0.6595712
racelen  0.7906728 1.0000000 0.3186467 0.6015968
cautions 0.1599884 0.3186467 1.0000000 0.3167884
leadchng 0.6595712 0.6015968 0.3167884 1.0000000
> R.lam <- eigen(R)$val
> R.e <- eigen(R)$vec
> round(R.lam,4)
[1] 2.5092 0.8919 0.4185 0.1805
> round(R.e,4)
        [,1]    [,2]    [,3]    [,4]
[1,] -0.5595  0.3367 -0.2034  0.7295
[2,] -0.5662  0.1170 -0.5160 -0.6321
[3,] -0.2903 -0.9338 -0.1106  0.1775
[4,] -0.5311  0.0311  0.8247 -0.1917
```

p.6.a. For the factor analytic model, with m = 1, compute estimates of **L** and **Ψ** based on the principal components method.

$\tilde{\mathbf{L}}$ = ______________________________ $\tilde{\mathbf{\Psi}}$ = ______________________________

p.6.b. What propotion of the standardized sample variance is due to the first factor?

Q.7. Q.5. A discriminant analysis is conducted to classify NHL and EPL players by Height and Weight. Random samples of $n_{NHL} = n_{EPL} = 100$ players to generate Fisher's discriminant function to classify players by league. The results for the 2 samples are given below.

```
        Xbar1    Xbar2     Diff      Sum Spooled          INV(Sp)
S111  73.3708  72.2699  1.1009 145.6407  5.6294  22.7423  0.3005 -0.0304
S112 202.4500 169.9500 32.5000 372.4000 22.7423 224.6843 -0.0304  0.0075
```

p.7.a. Compute $\hat{\mathbf{a}}' = \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)' S_{\text{pooled}}^{-1}$ and $\hat{m} = \frac{1}{2}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)' S_{\text{pooled}}^{-1}\left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right)$

$\hat{\mathbf{a}}' =$ ______________________________ $\hat{m} =$ ______________________________

p.7.b. The confusion matrix for the holdout samples (617 NHL players and 426 EPL players) is given below, based on the function generated for the training sample. Compute the estimate of the Expected actual error rate.

```
> (classtab <- table(league,classify))
      classify
league   1   2
     1 517 100
     2  95 331
```

$\hat{E}\{AER\} =$ ____

Have a Great Summer!