Thursday 6/11/2015 Final Examination 120 minutes

1. (40 pts.) Fifty bars of soap are manufactured in each of two ways ($n_1 = n_2 = 50$). Two characteristics, X_1 = lather and X_2 = mildness, are measured. The summary statistics for bars produced by methods 1 and 2 are

$$\bar{\boldsymbol{x}}_1 = \begin{bmatrix} 8.3 \\ 4.1 \end{bmatrix}, \qquad \boldsymbol{S}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}$$
$$\bar{\boldsymbol{x}}_2 = \begin{bmatrix} 10.2 \\ 3.9 \end{bmatrix}, \qquad \boldsymbol{S}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

In (a),(b),(c) below, assume the measurements follow bivariate normal distributions with common covariance of the mesurements of the two methods. In (d), do not assume the normality nor the common covariance of the data. Use the quantile notations in Figure 1 on the last page to construct the confidence intervals.

```
> xbar1 <- c(8.3, 4.1)
> xbar2 <- c(10.2, 3.9)
> S1 <- rbind(c(2,1), c(1,6))
> S2 <- rbind(c(2,1), c(1,4))
> n1 <- 50
> n2 <- 50
> p <- 2
> Sp <- ((n1-1)*S1 + (n2-1)*S2)/(n1+n2-2)
> xbar12 <- xbar1 - xbar2
> lambda <- eigen(Sp)$values
> vector <- eigen(Sp)$vectors
> Sp
     [,1] [,2]
[1,]    2    1
[2,]    1    5
> xbar12
[1] -1.9  0.2
> lambda
[1] 5.302776 1.697224
> vector
          [,1]       [,2]
[1,] 0.2897841 -0.9570920
[2,] 0.9570920  0.2897841
>
```

(a) Obtain 95% confidence region for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ (specify the center and the half-lengths of the major and the minor axes of the ellipse)

- Center: $(-1.9, 0.2)$

- half-lengths of axes: Let $n = n_1 + n_2 - 1 = 99$. If we let

$$c^2 = 98\frac{2}{97}F_{2,97}(0.95) = 2.02F_{2,97}(0.95)$$

then the half lengths are

$$\sqrt{5.302776}\sqrt{(\frac{1}{50}+\frac{1}{50})c^2}, \qquad \sqrt{1.697224}\sqrt{(\frac{1}{50}+\frac{1}{50})c^2}$$

(b) Obtain 95% simultaneous confidence intervals for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$

- $$\mu_{11} - \mu_{21} : -1.9 \pm \sqrt{2.02F_{2,97}(0.95)}\sqrt{(\frac{1}{50}+\frac{1}{50})2}$$
$$\mu_{12} - \mu_{22} : 0.2 \pm \sqrt{2.02F_{2,97}(0.95)}\sqrt{(\frac{1}{50}+\frac{1}{50})5}$$

(c) Obtain 95% Bonferroni intervals for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$

- $$\mu_{11} - \mu_{21} : -1.9 \pm t_{98}(0.975/2)\sqrt{(\frac{1}{50}+\frac{1}{50})2}$$
$$\mu_{12} - \mu_{22} : 0.2 \pm t_{98}(0.975/2)\sqrt{(\frac{1}{50}+\frac{1}{50})5}$$

(d) Obtain 95% large sample confidence intervals for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$

- since $n_1 = n_2$,

$$\mu_{11} - \mu_{21} : -1.9 \pm \sqrt{\chi^2_2(0.95)}\sqrt{(\frac{1}{50}+\frac{1}{50})2}$$
$$\mu_{12} - \mu_{22} : 0.2 \pm \sqrt{\chi^2_2(0.95)}\sqrt{(\frac{1}{50}+\frac{1}{50})5}$$

2. (10 pts.) State the Bonferroni method and prove the inequality of the method when $p = 3$.

- Let C_i denote the confidence interval for ith mean μ_i of p-dimensional vector $\boldsymbol{\mu}$ and let $\Pr(C_i) = 1 - \alpha$ for a given significance level α. The Bonferroni method is a method adjusting the one-at-a-time (individual) significance level α to α/p so that the simultaneous confidence intervals $C_1 \cap \cdots \cap C_p$ based on the individual significance level α/p satisfies

$$\Pr(C_1 \cap \cdots \cap C_p) \geq \alpha$$

- Proof when $p = 3$

$$\begin{aligned}\Pr(C_1 \cap C_2 \cap C_3) &= 1 - \Pr(C_1^c \cup C_2^c \cup C_3^c)\\ &\geq 1 - \Pr(C_1^c) - \Pr(C_2^c) - \Pr(C_3^c)\\ &= 1 - \frac{\alpha}{3} - \frac{\alpha}{3} - \frac{\alpha}{3} = 1 - \alpha.\end{aligned}$$

3. (40 pts.) Consider the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}$$

of a random vector $\boldsymbol{X}' = [X_1, X_2]$.

(a) Obtain the eigen-pairs of $\boldsymbol{\Sigma}$.

```
> S <- rbind(c(1,.4), c(.4,1))
> lambda <- eigen(S)$values
> vector <- eigen(S)$vector
> S
     [,1] [,2]
[1,]  1.0  0.4
[2,]  0.4  1.0
> lambda
[1] 1.4 0.6
> vector
          [,1]       [,2]
[1,] 0.7071068 -0.7071068
[2,] 0.7071068  0.7071068
```

(b) Obtain the two principal compoments Y_1, Y_2 from $\boldsymbol{\Sigma}$.

- $$\begin{aligned} Y_1 &= \boldsymbol{e}_1'\boldsymbol{X} = 0.707X_1 + 0.707X_2 \\ Y_2 &= \boldsymbol{e}_2'\boldsymbol{X} = -0.707X_1 + 0.707X_2 \end{aligned}$$

(c) Obtain the proportion of the total population variance explained by the first principal component.

```
> lambda[1]/sum(lambda)
[1] 0.7
```

(d) Obtain the correlations ρ_{Y_1,X_1} and ρ_{Y_1,X_2}

- $$\begin{aligned} \rho_{Y_1,X_1} &= \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = 0.707\sqrt{1.4} = 0.837 \\ \rho_{Y_1,X_2} &= \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = 0.707\sqrt{1.4} = 0.837 \end{aligned}$$

4. (20 pts.) Let

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

be a population correlation matrix where $0 < \rho < 1$. Obtain the proportion of the total variation explained by the first principal component of $\boldsymbol{R}$.

- Since $\boldsymbol{R} = \boldsymbol{I} + \rho(\boldsymbol{1}\boldsymbol{1}' - \boldsymbol{I})$, any eigenvector $\boldsymbol{x}$ of $\boldsymbol{R}$ satisfies $\boldsymbol{x} + \rho(3\bar{x}\boldsymbol{1} - \boldsymbol{x}) = \lambda\boldsymbol{x}$ where $\boldsymbol{I}$ is the 3×3 identity matrix and $\bar{x}$ is the average of the entries of $\boldsymbol{x}$. Thus

  $$(1 - \rho - \lambda)\boldsymbol{x} = -3\rho\bar{x}\boldsymbol{1}$$

  If $\boldsymbol{x}$ is parallel to $\boldsymbol{1}$, then $1 - \rho - \lambda = -3\rho$, that is $\lambda = 1 + 2\rho$. If $\boldsymbol{x}$ is not parallel to $\boldsymbol{1}$, then $\lambda = 1 - \rho$. Under the condition $\rho > 0$, the proportion of the total variation explained by the first principal component is

  $$\frac{1 + 2\rho}{3}$$

5. (20 pts.) Suppose that

$$\boldsymbol{R} = \begin{bmatrix} 1 & .63 & .45 \\ .63 & 1 & .35 \\ .45 & .35 & 1 \end{bmatrix}$$

is a correlation matrix for the $p = 3$ standardized random variables, Z_1, Z_2 and Z_3. $\boldsymbol{R}$ is generated by the one factor model

$$Z_1 = .9F_1 + \epsilon_1$$
$$Z_2 = .7F_1 + \epsilon_2$$
$$Z_3 = .5F_1 + \epsilon_3$$

where $\text{Var}(F_1) = 1, \text{Cov}(\boldsymbol{\epsilon}, F_1) = \boldsymbol{0}$.

(a) Compute communalities and specific variances.

- The factor model is $\boldsymbol{Z} = \boldsymbol{L}F_1 + \boldsymbol{\epsilon}$ where $\boldsymbol{L}' = [.9, .7, .5]$. Thus the communalities are $.81, .49, .25$ and

$$\boldsymbol{\Psi} = \boldsymbol{R} - \boldsymbol{L}\boldsymbol{L}' = \begin{bmatrix} .1 & & \\ & .3 & \\ & & .5 \end{bmatrix}$$

The specific variances are $.1, .3, .5$.

(b) Eigenvectors of $\boldsymbol{R}$ are $\boldsymbol{e}_1' = [.625, .593, .507]$, $\boldsymbol{e}_2' = [-.219, -.491, .843]$, $\boldsymbol{e}_3' = [.749, -.638, -.177]$. Eigenvalues of $\boldsymbol{R}$ are $\lambda_1 \geq \lambda_2 \geq \lambda_3 > 0$ whose eigen vectors are $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, and $\boldsymbol{e}_3$, respectively. Assuming an one factor model, compute the loading matrix $\boldsymbol{L}$ and the specific variance matrix $\boldsymbol{\Psi}$ based on the principal compoment solution for the factor model. Also compute the proportion of the total population variance which is explained by the estimated model.

```
> R <- cbind(c(1,.63,.45), c(.63,1,.35), c(.45,.35,1))
> e1 <- c(.625,.593,.507)
> l1 <- (R %*% e1)/e1      ## note that R %*% e1 = l1 %*% e1
> l1                       ## in principle the three entries of l1 are same
         [,1]
[1,] 1.962784
[2,] 1.963238
[3,] 1.964103
> l1 <- mean(l1)           ## so just take the mean of l1
> l1
[1] 1.963375

> Lhat <- cbind(sqrt(l1)*e1)  ## loading matrix
> Lhat %*% t(Lhat)
          [,1]      [,2]      [,3]
[1,] 0.7669433 0.7276758 0.6221444
[2,] 0.7276758 0.6904188 0.5902906
[3,] 0.6221444 0.5902906 0.5046835
> psihat <- diag(R - Lhat %*% t(Lhat))  ## specific
> psihat
[1] 0.2330567 0.3095812 0.4953165 ## specific variances
```

```
>
> l1/3
[1] 0.6544583  ## the proportion
>
```

6. (20 pts.) Recall that the conditional mean vector of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is

$$\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$$

where

$$\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim N(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix})$$

(a) Using the result above, derive the formula of the factor score of the factor model $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\Psi}$ where $\boldsymbol{\mu} = E[\boldsymbol{X}]$, $\boldsymbol{L}$ is the loading matrix, $\boldsymbol{F}$ is the common factor, and $\boldsymbol{\Psi}$ is the matrix of specific variances.

- $$\text{Cov}(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{F} \end{bmatrix}) = \begin{bmatrix} \text{Cov}(\boldsymbol{X}, \boldsymbol{X}) & \text{Cov}(\boldsymbol{X}, \boldsymbol{F}) \\ \text{Cov}(\boldsymbol{F}, \boldsymbol{X}) & \text{Cov}(\boldsymbol{F}, \boldsymbol{F}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi} & \boldsymbol{L} \\ \boldsymbol{L}' & \boldsymbol{I} \end{bmatrix}$$

  implies

$$\begin{aligned} E[\boldsymbol{F}|\boldsymbol{x}] &= \boldsymbol{L}'(\boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \\ \Longrightarrow \hat{\boldsymbol{f}}^R &= \hat{\boldsymbol{L}}'(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\boldsymbol{\Psi}})^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}) \end{aligned}$$

(b) Suppose a maximum likelihood solution from a correlation matrix $\boldsymbol{R}$ gave the following estimated loadings and specific variances:

$$\hat{\boldsymbol{L}} = \begin{bmatrix} .1 & .2 \\ .3 & .4 \\ .5 & .6 \end{bmatrix}, \qquad \hat{\boldsymbol{\Psi}} = \begin{bmatrix} .1 & 0 & 0 \\ 0 & .2 & 0 \\ 0 & 0 & .3 \end{bmatrix}$$

Using the formula obtained in part (a), compute the factor score for the standardized observation $\boldsymbol{z}' = [7, 8, 9]$.

```
> Lhat <- rbind(c(.1,.2), c(.3,.4), c(.5,.6))
> Psi <- diag(c(.1, .2, .3))
> z <- c(7,8,9)
> t(Lhat) %*% solve(Lhat %*% t(Lhat) + Psi) %*% z
          [,1]
[1,]  6.004112
[2,] 10.938999
>
```

7. (30 pts.) Consider two bivariate normal populations π_1 and π_2. Assume the distributions of π_1 and π_2 are

$$\pi_1 \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}), \qquad \pi_2 \sim N(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix})$$

and the prior probabilities are $p_1 = p_2 = \frac{1}{2}$.

(a) Compute the posterior probability of $\boldsymbol{x}' = [1, 0]$ being classified to π_1.

```
> f1 <- function (x) dnorm(x[1])*dnorm(x[2])
> f2 <- function (x) dnorm(x[1], mean=1)*dnorm(x[2], mean=1)
> p1 <- p2 <- 1/2
> posterior1 <- function (x) p1*f1(x)/(p1*f1(x) + p2*f2(x))
> posterior1(c(1,0))
[1] 0.5
>
```

(b) Find the classification rule (R_1, R_2) based on the posterior probability.

- $$\begin{aligned}
\Pr(\boldsymbol{x} \sim \pi_1 | \boldsymbol{x}) &= \frac{p_1 f_1(\boldsymbol{x})}{p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})} \\
&= \frac{1}{1 + (\exp -\frac{1}{2}(x_1 - 1)^2 - \frac{1}{2}(x_2 - 1)^2)/(\exp -\frac{1}{2}x_1^2 - \frac{1}{2}x_2^2)} \\
&= \frac{1}{1 + \exp(x_1 + x_2 - 1)} \\
&\geq \frac{1}{2} \\
\Longrightarrow x_1 + x_2 - 1 &\leq 0
\end{aligned}$$

  Thus

  $$R_1 : x_1 + x_2 - 1 \leq 0, \qquad R_2 : x_1 + x_2 - 1 > 0$$

(c) Based on the classification rule in (b), obtain the actual error rate (the probability of misclassification).

- Let $y = \binom{1}{1}' \boldsymbol{x}$. Then if $\boldsymbol{x} \sim \pi_1$ then $y \sim N(0, 2)$ and if $\boldsymbol{x} \sim \pi_2$ then $y \sim N(2, 2)$.

  $$\begin{aligned}
\Pr(\boldsymbol{x} \in R_2 | \boldsymbol{x} \sim \pi_1) &= \Pr(y > 1 | \mu = 0, \sigma^2 = 2) = 1 - \Phi(\frac{1}{\sqrt{2}}) = \Phi(-\frac{1}{\sqrt{2}}) \\
\Pr(\boldsymbol{x} \in R_1 | \boldsymbol{x} \sim \pi_2) &= \Pr(y \leq 1 | \mu = 2, \sigma^2 = 2) = \Phi(-\frac{1}{\sqrt{2}}) \\
AER &= \Phi(-\frac{1}{\sqrt{2}}) = 0.24
\end{aligned}$$

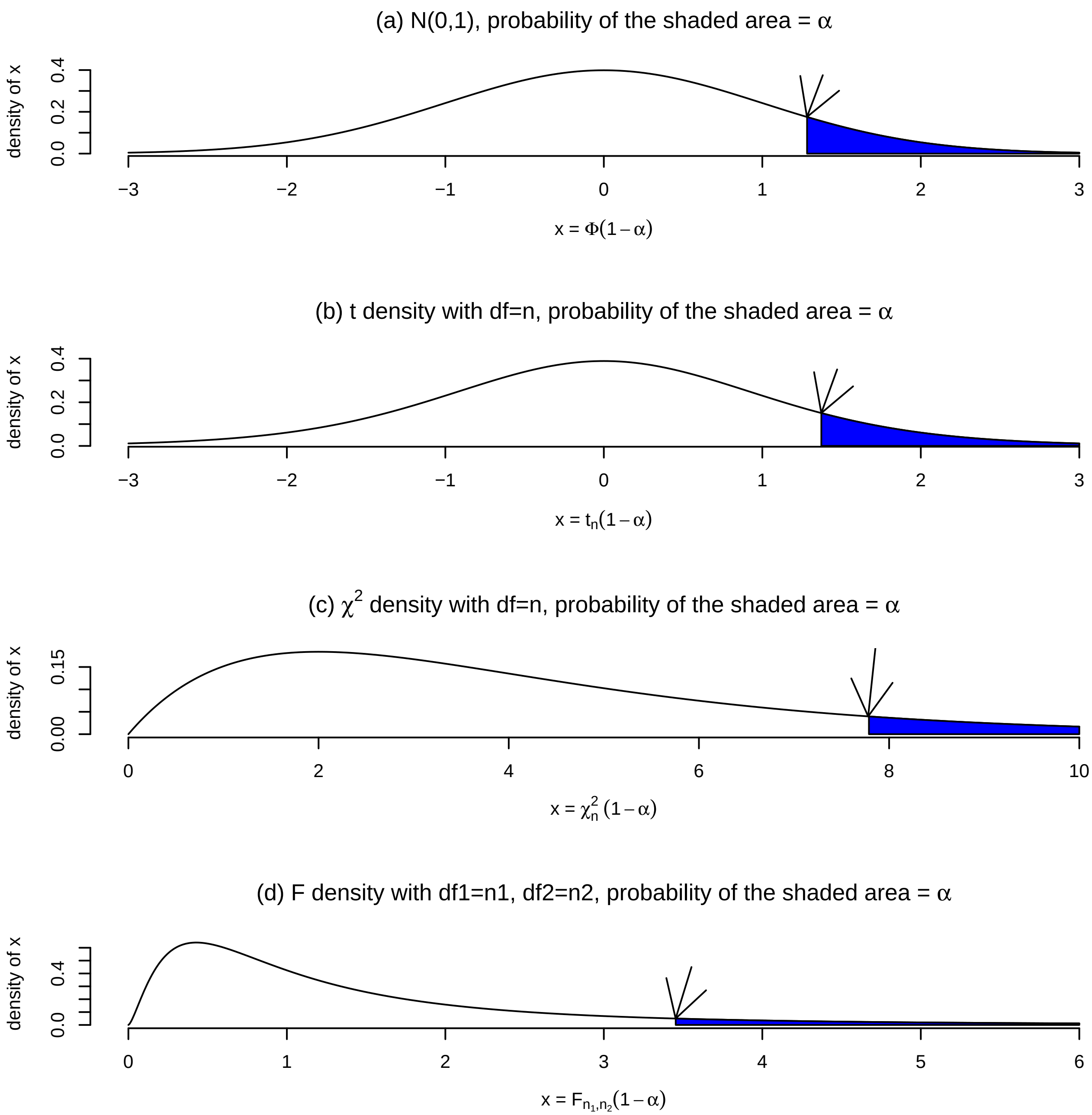

Figure 1: $100(1-\alpha)\%$ percentiles are (a) $\Phi(1-\alpha)$, (b) $t_n(1-\alpha)$, (c) $\chi^2_n(1-\alpha)$, (d) $F_{n_1,n_2}(1-\alpha)$.