UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER 2014 EXAMINATIONS

STA437H1F / STA2005H1F

Duration - 3 hours

Examination Aids: A nonprogrammable calculator

Last name: ______________________________

First name: ______________ Middle name: ______________

Student number: ______________________________

Instruction

1. If a question asks you to so some calculations or derivations, you must show your work to receive full credit. Simplify expressions as much as you can.
2. If you do not have enough space, use the other side and refer it correctly.
3. Some problems could be wrong. If it happens, correct questions by yourself after arguing why the problem is wrong.
4. Please do not pull the pages apart unless necessary. If you do so, print your name and sign all the pages.
5. You can use either pen or pencil for the test. **But please be aware that you are not allowed to dispute any credit after the test is returned if you use pencil.**
6. You are allowed to use any theorem covered in class as well as any statement in other questions in this exam. For example, in question x (y), you can use the result of question x (z) without a proof.
7. There are 3 problems on 11 sheets, and the total is 100 marks. Please hand in your exam paper and books when the time is up.
8. Some useful **information** are on page 11.

PLEASE HAND IN THIS QUESTION PAPER

Do NOT enter ANY answers on this yellow question paper. It will be destroyed after the exam. Answers on it will NOT be marked.

Problem 1 (50 marks). A zoologist collected three measurements of $n = 25$ lizards. Measurements are the weight (Mass), snout-vent length (SVL) and hind limb span (HLS), say three random variables $X = (X_1, X_2, X_3)$. The summary statistics are

$$\bar{\mathbf{x}} = \begin{pmatrix} 8.69 \\ 68.40 \\ 129.32 \end{pmatrix}, S = \begin{pmatrix} 7.187 & 20.700 & 33.537 \\ 20.700 & 63.771 & 102.085 \\ 33.537 & 102.085 & 185.831 \end{pmatrix} R = \begin{pmatrix} 1.000 & 0.967 & 0.918 \\ 0.967 & 1.000 & 0.938 \\ 0.918 & 0.938 & 1.000 \end{pmatrix}$$

$$S^{-1} = \begin{pmatrix} 2.172 & -0.643 & -0.039 \\ -0.643 & 0.320 & -0.060 \\ -0.039 & -0.060 & 0.045 \end{pmatrix} R^{-1} = \begin{pmatrix} 15.617 & -13.760 & -1.419 \\ -13.760 & 20.423 & -6.525 \\ -1.419 & -6.525 & 8.421 \end{pmatrix}.$$

Eigen values and vectors of $S = U_S \Lambda_S U_S^\top$ and $R = U_R \Lambda_R U_R^\top$ are

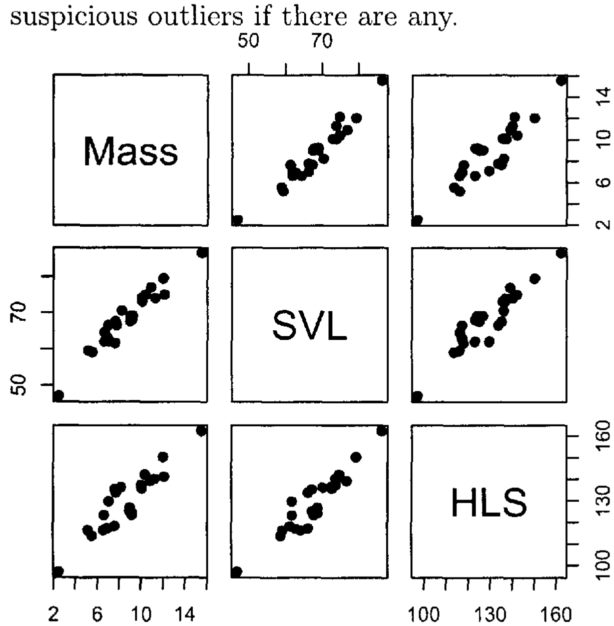
$$\Lambda_S = \begin{pmatrix} 250.123 & 0 & 0 \\ 0 & 6.244 & 0 \\ 0 & 0 & 0.421 \end{pmatrix}, U_S = \begin{pmatrix} -0.160 & -0.252 & 0.954 \\ -0.488 & -0.820 & -0.299 \\ -0.858 & 0.513 & -0.008 \end{pmatrix},$$

$$\Lambda_R = \begin{pmatrix} 2.882 & 0 & 0 \\ 0 & 0.088 & 0 \\ 0 & 0 & 0.031 \end{pmatrix}, U_R = \begin{pmatrix} -0.578 & -0.543 & 0.609 \\ -0.582 & -0.249 & -0.774 \\ -0.572 & 0.802 & 0.172 \end{pmatrix}.$$

(a) [5 marks] Describe a general procedure of detecting outliers.

CONT'D.

(b) [3 marks] Scatter plot of the data is given below. Comment on possible outliers. Circle suspicious outliers if there are any.



(c) [10 marks] The followings are four most anomalous observations. Pick one of the most suspicious observation and check whether or not the chosen one is an outlier.

ID	Mass	SVL	HLS
5	7.063	62.0	129.5
8	2.447	47.0	97.0
9	15.493	86.5	162.0
24	6.978	66.5	117.0

CONT'D.

(d) [4 marks] Express three principal components of S and compute the proportion of the variance of each component.

(e) [4 marks] Express three principal components of R and compute the proportion of the variance of each component.

(f) [4 marks] Discuss which one is more appropriate to use in principal component analysis between variance S and correlation R.

CONT'D.

(g) [5 marks] Describe in general how to assess equal correlation assumption.

(h) [5 marks] Discuss whether or not the data support equal correlation hypothesis.

CONTD

(i) [10 marks] Let $Y = X_1 - \bar{\mathbf{x}}_1\mathbf{1}$, $Z_1 = X_2 - \bar{\mathbf{x}}_2\mathbf{1}$, $Z_2 = X_3 - \bar{\mathbf{x}}_3\mathbf{1}$ and $Z = (Z_1\ Z_2)$. Fit the linear regression model $Y = \beta_1 Z_1 + \beta_2 Z_2 + E$ where E is error vector having mean zero and constant variance.

Problem 2 (35 marks). A zoologist collected weights (Mass) and snout-vent length (SVL) of two species of lizards, that is, Cnemidophorus (X_1) and Sceloporus (X_2). In the following data analysis, log transformed data is used. Descriptive statistics are as follows.

$$n_1 = 20, \bar{\mathbf{x}}_1 = \begin{pmatrix} 2.240 \\ 4.394 \end{pmatrix}, S_1 = \begin{pmatrix} 0.353 & 0.094 \\ 0.094 & 0.026 \end{pmatrix} n_2 = 40, \bar{\mathbf{x}}_2 = \begin{pmatrix} 2.368 \\ 4.308 \end{pmatrix}, S_2 = \begin{pmatrix} 0.507 & 0.145 \\ 0.145 & 0.043 \end{pmatrix}$$

(a) [2 marks] Compute correlation coefficients in two groups, that is, $\mathrm{cor}(x_{i1}, x_{i2})$ for $i = 1, 2$.

(b) [2 marks] Find the pooled variance matrix.

(c) [4 marks] Assess $H : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ under the assumption $\Sigma_1 = \Sigma_2$.

(d) [4 marks] Assess $H : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ under the assumption $\Sigma_1 \neq \Sigma_2$.

(e) [7 marks] Find a rectangular shape 95%-confidence region for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

(f) [6 marks] Derive the likelihood ratio test statistic for $H : \Sigma_1 = \Sigma_2$.

(g) [10 marks] Assess $H : \boldsymbol{\mu} = (2.25, 4.5)$ under the assumption $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$ and $\Sigma_1 = \Sigma_2$.

Problem 3 (15 marks). Let X be $N_3(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = (1, 2, -3)^\top$ and $\Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$

(a) [5 marks] Find the distribution of $3X_1 + 2X_1 + X_3$

(b) [5 marks] Find the conditional distribution of X_2 given $X_1 = 3$, $X_3 = 1$.

(c) [5 marks] Find a so that X_1 and $aX_1 + 2X_2 + X_3$ are independent.

CONT'D.

Useful Information

1. Let $X \sim N_p(\mu, \Sigma)$. (a) The density and moment generating functions of X are $|2\pi\Sigma|^{-1/2}\exp(-(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)/2)$ and $\exp(\mathbf{t}^\top\mu) + \mathbf{t}^\top\Sigma\mathbf{t}/2)$.
(b) When $X = (X_1^\top, X_2^\top)^\top$, the conditional distribution of X_1 given X_2 is $N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$.
2. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots$ be i.i.d. sample with $\mu = \mathbb{E}(\mathbf{x}_i) \in \mathbb{R}^p$ and $\Sigma = \mathbb{V}\mathrm{ar}(\mathbf{x}_i) \in \mathbb{R}^{p\times p}$.
(a) $\bar{\mathbf{x}} = (\mathbf{x}_1 + \cdots + \mathbf{x}_n)/n$, $S = \frac{1}{n-1}\sum_{i=1}^n(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$
(b) T^2-statistic is defined by $T^2 = n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu)$.
(c) $\max_{\mathbf{a}}[n(\mathbf{a}^\top(\bar{\mathbf{x}} - \mu))^2/(\mathbf{a}^\top S\mathbf{a})] = n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu)$.
(d) Law of large numbers: $\bar{\mathbf{x}} \to \mu$ almost surely; Central limit theorem: $\sqrt{n}(\bar{\mathbf{x}} - \mu) \longrightarrow N(O, \Sigma)$ in distribution.
(e) If $\mathbf{x}_i \sim N_p(\mu, \Sigma)$, then $T^2 = n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \sim F(p, n-p)(n-1)p/(n-p)$ and for $\mathbf{a} \in \mathbb{R}^p$, $\mathbf{a}^\top(\bar{\mathbf{x}} - \mu)/\sqrt{\mathbf{a}^\top S\mathbf{a}/n} \sim t(n-1)$.
3. Continuous mapping theorem: If $\sqrt{n}(\bar{\mathbf{x}} - \mu) \to N(O, \Sigma)$ in distribution and g is continuous, then $\sqrt{n}(g(\bar{\mathbf{x}}) - g(\mu)) \to g'(\mu)N(O, \Sigma)$ in distribution.
4. The density and moment generating functions of $A \sim \mathrm{Wishart}_p(n, \Sigma)$ are $\mathrm{pdf}_{\mathbf{A}}(\mathbf{A}) = |\mathbf{A}|^{(m-p-1)/2}\exp(-\mathrm{tr}(\Sigma^{-1}\mathbf{A})/2)/[2^{np/2}|\Sigma|^{n/2}\Gamma_p(n/2)]$ and $\mathrm{mgf}_{\mathbf{A}}(U) = |I_p - 2U\Sigma|^{-n/2}$ for some $U \in \mathbb{R}^{p\times p}$ around O.
5. Some useful quantile tables from R package.

$t_\gamma(n)$	$\gamma = 0.9$	0.95	0.975	0.9875
$n = 4$	1.533	2.132	2.776	3.495
$n = 5$	1.476	2.015	2.571	3.163
$n = 58$	1.296	1.672	2.002	2.301
$n = 59$	1.296	1.671	2.001	2.300
$n = 60$	1.296	1.671	2.000	2.299
$n = \infty$	1.282	1.645	1.960	2.241

$\chi^2_\gamma(n)$	$\gamma = 0.9$	0.95	0.975	0.99
$n = 2$	4.605	5.991	7.378	9.210
$n = 3$	6.251	7.815	9.348	11.345
$n = 58$	72.160	76.778	80.936	85.950
$n = 59$	73.279	77.931	82.117	87.166
$n = 60$	74.397	79.082	83.298	88.379

$F_\gamma(m, n)$	$\gamma = 0.9$	0.95	0.975
$m = 2, n = 3$	5.462	9.552	16.044
$m = 2, n = 58$	2.396	3.156	3.934
$m = 2, n = 59$	2.395	3.153	3.929
$m = 2, n = 60$	2.393	3.15	3.925
$m = 3, n = 58$	2.181	2.764	3.351
$m = 3, n = 59$	2.179	2.761	3.347
$m = 3, n = 60$	2.177	2.758	3.343