Midterm - CSC/DSC 265/465 - Intermediate Statistical and Computational Methods - 03/20/2018 - WITH SOLUTIONS

NAME: _____

PLEASE NOTE that all students will do a total of 5 questions.

Undergraduate students do questions 1-5.

Graduate students do questions 1-4 and 6.

The exam will last one hour and ten minutes. You are allowed up to five aid sheets on standard 8.5×11 inch paper (both sides) and a calculator. Answer the questions in the space provided. Use the back of the sheet if needed (please indicate if you have done this). You are encouraged to read each question completely before starting. Critical values for the t distributions are given in tabular form on the final page of this exam sheet. No other ctitical values will be needed.

Q1: An ANOVA model is analyzed based on data for 4 treatments, of which each have n_j observations. An observation from treatment $j = 1, \ldots, 4$ has distribution $N(\mu_j, \sigma^2)$, the variance being assumed constant. Observations are independent. The treatment means, standard deviations and sample sizes are given in a table below. The sum of squares (SS), mean sum of squares (MSS) and degrees of freedom for treatment and error sources of variation are given in the subsequent table.

Using a Bonferroni multiple comparison procedure, with a family-wise error rate of $\alpha_{FWE} = 0.15$, can we conclude that μ_4 is the minimum treatment mean?

Treatment j	$\bar{X}_j$	S_j	n_j
1	10.03	1.39	7
2	10.24	1.31	7
3	13.22	1.21	7
4	6.92	2.06	7

Source of Variation	DF	SS	MSS
Treatment	3	139.50	46.50
Error	24	56.15	2.34

SOLUTION We need $m = 3$ comparisons, to compare $\mu_4 - \mu_i$, $i = 1, 2, 3$. Since $n_i = 7$ for $i = 1, 2, 3, 4$ and $n = 28$, the CIs take form

$$\begin{aligned} CI &= \bar{X}_i - \bar{X}_j \pm t_{n-k,\alpha_{FWE}/(m2)} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \\ &= \bar{X}_i - \bar{X}_j \pm t_{24,0.15/6} \sqrt{2.34\left(\frac{1}{7} + \frac{1}{7}\right)} \\ &= \bar{X}_i - \bar{X}_j \pm 2.064 \sqrt{2.34\left(\frac{1}{7} + \frac{1}{7}\right)} \\ &= \bar{X}_i - \bar{X}_j \pm 1.684. \end{aligned}$$

The CIs are given in the following table. We can conclude with confidence $1 - \alpha_{FWE} = 0.85$ that μ_4 is the smallest mean, since $\mu_4 - \mu_j < 0$ for $j = 1, 2, 3$ within each comparison.

Comparison	Estimate	ME	LB	UB
$\mu_4 - \mu_1$	-3.11	1.69	-4.80	-1.42
$\mu_4 - \mu_2$	-3.32	1.69	-5.01	-1.63
$\mu_4 - \mu_3$	-6.30	1.69	-7.99	-4.61

Q2: A client hires a consulting firm to conduct a study of two types of mutual funds (we'll call them simply Type A and Type B). It uses a simple regression model

$$Y = e^{\beta_0+\beta_1 X+\epsilon}$$

where $X = 1$ for a Type A mutual fund, and $X = 0$ otherwise; $\epsilon \sim N(0, \sigma^2)$; and Y is the value of an original investment of \$1 after a year (that is, if $Y = 1.05$, the yearly rate of return is 5%). The model is first log-transformed, giving

$$\log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i, \;\; i = 1, \ldots, n. \tag{1}$$

A random sample of $n = 62$ paired observations (Y_i, X_i), $i = 1, \ldots, 62$ is collected. A simple least squares regression model is used to fit the model (1), producing the following coefficient table:

Coefficient	Estimate	Standard Error	t-value	Pr(>|t|)
$\hat{\beta}_0$	0.0745	0.0040	18.8376	2.43×10^{-34}
$\hat{\beta}_1$	0.0333	0.0056	5.9514	4.13×10^{-8}

The consultant believes Type A mutual funds have a higher average yield, but the client currently purchases mutual funds of Type B, and there would be a significant cost to switching to Type A. Therefore, the consultant will only recommend switching to Type A if there is significant statistical evidence that $\beta_1 > 0.015$ (approximately, that the rate of return of Type A mutual finds exceeds Type B mutual funds by more than 1.5%). Using a significance level of $\alpha = 0.05$, can the consultant recommend switching?

SOLUTION The hypotheses are

$$H_o : \beta_1 \leq 0.015 \text{ against } H_a : \beta_1 > 0.015.$$

The appropriate t-statistic is

$$T = \frac{\hat{\beta}_1 - 0.015}{SE_{\hat{\beta}_1}} = \frac{0.0333 - 0.015}{0.0056} \approx 3.268.$$

Since$T > t_{60,0.05} = 1.671$ we do not reject the conjecture at significance level $\alpha = 0.05$.

Q3: Suppose we observe a normally distributed random variable $X \sim N(\mu, \sigma)$. Then X has density function

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in (-\infty, \infty).$$

Assume σ is known, and that μ has a prior density $\pi(\mu)$:

$$\mu \sim N(\mu_0, \sigma_0),$$

for some fixed μ_0, σ_0. Show that the normal prior density is a conjugate density for μ, that is, that the posterior density for μ given X is also normal. Give this density precisely.

SOLUTION Recall that to evaluate a posterior density of a parameter θ given data x it is often easiest to first express it as

$$\pi(\theta \mid x) = Kg(\theta)$$

where K is a constant that does not depend on θ, and then normalize $g(\theta)$. This means we don't need to actually evaluate K. In this case we have

$$\pi(\mu \mid x) \propto f(x \mid \mu)\pi(\mu)$$

where $x \mid \mu \sim N(\mu, \sigma)$ and $\mu \sim N(\mu_0, \sigma_0)$. This means

$$\pi(\mu \mid x) = Ke^{-\frac{1}{2}Q_1} e^{-\frac{1}{2}Q_0}$$

where

$$Q_1 = \frac{(x-\mu)^2}{\sigma^2}, \quad Q_0 = \frac{(\mu-\mu_0)^2}{\sigma_0^2}.$$

We then have

$$Q_1 + Q_0 = \mu^2 \left[\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right] - 2\mu \left[\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right] + \left[\frac{x^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2}\right].$$

This means

$$\pi(\mu \mid x) = Ke^{-\frac{1}{2}\left\{\mu^2\left[\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right] - 2\mu\left[\frac{x}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right]\right\}}$$

where K does not depend on μ, and that $\pi(\theta \mid x) \sim N(\mu_{post}, \sigma^2_{post})$ is a normal density function with mean and variance

$$\begin{aligned} \mu_{post} &= \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}, \\ \sigma^2_{post} &= \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}. \end{aligned}$$

Q4: Suppose we observe survival times $21, 25+, 25, 27+, 34, 34, 35$. Recall that the symbol '+' denotes a right-censored observation. Construct and sketch a Kaplan-Meier estimate for the survival function.

SOLUTION For survival times $21, 25+, 25, 27+, 34, 34, 35$ we have table:

i	t_i	d_i	$r(t_i)$	$\hat{p}_i$
0	0	0	7	(7-0)/7 = 1
1	21	1	7	(7-1)/7 = 6/7
2	25	1	6	(6-1)/6 = 5/6
3	27	0	4	(4-0)/4 = 1
4	34	2	3	(3-2)/3 = 1/3
5	35	1	1	(1-1)/1 = 0

Then plot the cumulative products

$$\hat{p}_0, \hat{p}_0\hat{p}_1, \hat{p}_0\hat{p}_1\hat{p}_2, \ldots, \hat{p}_0\hat{p}_1 \times \cdots \times \hat{p}_5 = 1, 6/7, 5/7, 5/7, 5/21, 0$$

at times

$$t_0, \ldots, t_5 = 0, 21, 25, 27, 34, 35.$$

Note that '+' indicates the position of a censored observation. See Figure 1.

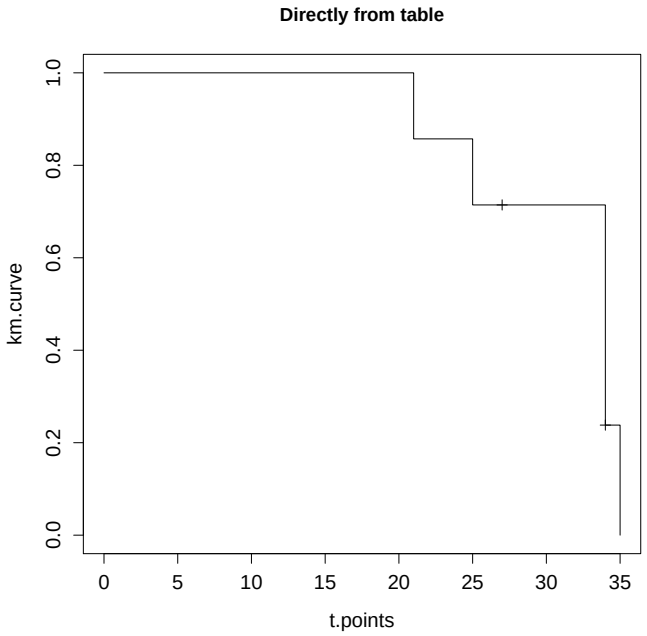

Figure 1: Kaplan-Meier estimate of survival function for Question 4

Q5 [Undergraduate Students Only]: Consider the case of linear regression through the origin:

$$Y_i = \beta X_i + \epsilon_i, \;\; i = 1, \ldots, n$$

where $\epsilon_i \sim N(0, \sigma^2)$ are iid error terms, and $X_1, \ldots, X_n$ are fixed predictor terms. Write explicitly the error sum of squares SSE for this model, where $\hat{\beta}$ is an estimate of β. After verifying that SSE is a second order polynomial in $\hat{\beta}$, determine the least squares estimate of β directly in terms of the observation (X_i, Y_i), $i = 1, \ldots, n$.

SOLUTION We have

$$SSE = \sum_{i=1}(Y_i - \hat{\beta} X_i)^2 = \left[\sum_{i=1} Y_i^2\right] - 2\hat{\beta}\left[\sum_{i=1} X_i Y_i\right] + \hat{\beta}^2\left[\sum_{i=1} X_i^2\right].$$

The minimum is directly given as

$$\hat{\beta} = \frac{\sum_{i=1} X_i Y_i}{\sum_{i=1} X_i^2}.$$

Q6 [Graduate Students Only]: Suppose we are given the simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{2}$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$. Suppose X is then interpreted itself as a random outcome with distribution $X \sim N(0,1)$, which is independent of ϵ. Derive the correlation coefficient ρ_X of (X, Y), where

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$$

SOLUTION The correlation is

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{var(X)var(Y)}}.$$

We have, given what we know of ϵ, X and Y,

$$\begin{aligned} \mu_X &= 0 \\ \mu_Y &= \beta_0 \\ var(X) &= 1 \\ var(Y) &= var(\beta_1 X) + var(\epsilon) = \beta_1^2 + \sigma_\epsilon^2 \\ E[(X-\mu_X)(Y-\mu_Y)] &= E[X(\beta_1 X + \epsilon)] = \beta_1 E[X^2] + E[X\epsilon] = \beta_1 + 0. \end{aligned}$$

So,

$$\rho_{XY} = \frac{\beta_1}{\sqrt{\beta_1^2 + \sigma_\epsilon^2}}.$$

Table 1: Critical values for the t distribution with ν degrees of freedom.

df $= \nu$	$t_{\nu,0.05}$	$t_{\nu,0.025}$
20	1.725	2.086
21	1.721	2.080
22	1.717	2.074
23	1.714	2.069
24	1.711	2.064
25	1.708	2.060
26	1.706	2.056
27	1.703	2.052
28	1.701	2.048
29	1.699	2.045
30	1.697	2.042
35	1.690	2.030
40	1.684	2.021
45	1.679	2.014
50	1.676	2.009
55	1.673	2.004
60	1.671	2.000
65	1.669	1.997
70	1.667	1.994
75	1.665	1.992
80	1.664	1.990
85	1.663	1.988
90	1.662	1.987
95	1.661	1.985
100	1.660	1.984