Uva Wellassa University of Sri Lanka
Faculty of Science and Technology
Department of Science and Technology
300 Level 2nd Semester Examination – Dec.2018/ Jan. 2019
SCT 333-1 Data Handling and Statistics

Instructions to candidates.
Duration: Two (02) hours
Number of questions: Four (04) esssay questions
Mark allocation: 100
Answer All Questions.

1.

a. Define the following terms that are used in hypothesis testing and explain how they relate to Type I and Type II errors.
   i. The significance level of the test. (3 mark)
   ii. The power of the test. (3 mark)

b. A soft drinks machine at a fast food outlet is regulated so that the amount of drink dispensed is normally distributed with standard deviation 15 ml. It is intended that the mean amount dispensed should be 200 ml. To test whether the machine is working correctly, it is checked periodically by taking a random sample of 9 drinks and calculating the mean amount dispensed. The machine is considered to be working satisfactorily if this mean lies between 191 ml and 209 ml.

   i. Write appropriate null and alternative hypotheses for this testing procedure. (3 mark)
   ii. State the distribution of the sample mean $\bar{x}$, if the null hypothesis is true. (3 mark)
   iii. Calculate the significance level of the test. If you were performing the test at the 5% level of significance, what would you conclude? (4 mark)
   iv. Calculate the power of the test if the true mean amount dispensed is 205 ml. Without performing any further calculations, state how you would expect this value to compare with the power of the test if the true mean amount dispensed is 210 ml, justifying your answer. (5 mark)

2. In Sri Lanka, all patients surviving a stroke are supposed to have their cholesterol levels measured soon after their stroke and regularly thereafter. A sample of medical records of men and women who had suffered a stroke was examined to determine whether there was a difference between the sexes in the proportions of stroke survivors who had a recently recorded cholesterol measurement. The following data were obtained.

	Cholesterol Level Recorded	
Sex	No	Yes
Female	109	22
Male	97	77

a. Perform a suitable test of the null hypothesis that there is no association between an individual's sex and the chance of he or she having a recently recorded cholesterol measurement. (7 mark)
b. Compute and interpret an approximate 95% confidence interval for the difference between the proportions of females and males having a recently recorded cholesterol measurement. (7 mark)

92

3. In an ecological study of 50 bears, regarded as a random sample, were temporarily anaesthetised, and their bodies were measured and weighed. One goal of the study was to estimate the weight of a bear based on other measurements. This would be used because in the forest it is easier to measure the length of a bear, for example, than it is to weigh it.

   The following variables were recorded: Sex (M-Male or F-Female); Age (age, in months); Head (length of the head, in inches); Head (width of the head, in inches); Neck (neck girth, distance around the neck, in inches); Chest (chest girth, in inches); Weight (weight, in pounds).

   The researchers wished to build a multiple linear regression model to predict the unknown weight of a bear using fitted model. Relavent data set is given below.
   (Further, data are stored in a Minitab work sheet named "Bear.mtw" given on your computer.)

ID	Sex	Age	Head_L	Head_W	Neck	Chest	Weight
41	F	23	12.5	5	20.5	38	142
48	M	81	15.5	8	31	54	416
69	M	25	16	8	32	52	432
83	M	117	15.5	7.5	32	54.5	476
238	M	70	15	6.5	28	45	334
274	F	57	13.5	7	20	38	204
518	M	45	13.5	7	24	39	204
520	F	9	9	4.5	12	19	26
522	M	21	13	6	19	30	120
525	M	177	16	9.5	30	48	436
527	F	57	12.5	5	19	32	125
529	F	81	13	5	20	33	132
531	M	21	13	5	17	28	90
533	M	9	10	4	13	23	40
535	M	45	16	6	24	42	220
538	M	9	10	4	13.5	23	46
541	M	33	13.5	6	22	34	154
543	F	57	13	5.5	17.5	31	116
545	F	45	13	6.5	21	34.5	182
547	M	21	14.5	5.5	20	34	150
549	M	10	9.5	4.5	16	26	65
551	F	82	13.5	6.5	28	48	356
553	F	70	14.5	6.5	26	48	316
555	M	10	11	5	17	29	94
557	M	10	11.5	5	17	29.5	86
560	M	34	13	7	21	35	150
562	M	34	16.5	6.5	27	44.5	270
564	M	34	14	5.5	24	39	202
568	F	58	13.5	6.5	21.5	40	202
570	M	58	15.5	7	28	50	365

572	M	11	11.5	6	16.5	31	79
576	M	23	12	6.5	19	38	148
579	M	70	15.5	7	28	55	446
581	F	11	9	5	15	27	62
583	F	83	14.5	7	23	44	236
585	M	35	13.5	8.5	23	44	212
598	M	35	14.5	6.5	26	41	278
626	M	16	13.5	6.5	22	36	190
630	M	24	13	6	19.5	31	140
644	M	70	13	5.5	19.5	35	162
648	M	19	13	5.5	19.5	32.5	126
659	M	75	17	8.5	30.5	48.5	368
661	M	36	12	5.5	18	32	116
663	F	87	14	7	21	37	160
665	M	41	13	6.5	20.5	36.5	154
670	M	71	16	7.5	28	45	316
673	F	36	13.5	5.5	19.5	35	158
675	F	23	12.5	5.5	19	32	120
679	M	81	15.5	7.5	25.5	43	324
681	M	56	14.5	7	22	38	196

a. Examine data carefully and state preliminary conclusions. (7 mark)

b. Use these data to develop suitable fitted equations for predecting weight of a bear in terms of the predictor variables using appropriate method. (8 mark)

c. Carry out a test at the 5% level of the significance of the variable Head_L in the regression, stating the hypothesis that you are testing. Calculate also a 90% confidence interval for the coefficient of Chest. (5 mark)

d. Write a brief report on your findings (it should inluded all relevent hypothesis with ANOVA, model checking with appropriate test and validation and necessary and relevant analysis). (15 mark)

4. In an ecological study, the presence or absence of an important plant species (used as an indicator of soil health) was recorded in a random selection of woodland sites across the UK, together with three categorical environmental variables –soil type (clay loam, silt loam, sandy loam), annual rainfall (low, medium, high), and altitude (low, high). The numbers of sites for each combination of the three environmental variables, and the numbers of sites with and without the indicator plant species for each combination, are shown below.

Altitude		Low		High	
Indicator Species		Present	Absent	Present	Absent
Soil Type	**Rainfall**				
Clay Loam	Low	12	4	5	10
	Medium	12	6	7	9
	High	11	5	6	11
Silt Loam	Low	10	7	8	11

	Medium	11	9	6	8
	High	7	5	9	13
Sandy Loam	Low	6	12	10	8
	Medium	4	13	11	7
	High	2	19	14	3

A series of log-linear models has been fitted in an attempt to identify how the environmental variables influence the health of the woodland soils, including different factors which are mention in the table.

a. What is the suitable model to describe the above situation (describe each and every parameter in the model)? (5 mark)
b. Construct the analysis of variance (ANOVA). What is your decision based on ANOVA? (Check validity of all possible hypotheses with suitable mean comparison) (10 mark)
c. Write a brief report on the results of the experiment in non-statistical terms. (15 mark)