Tutorial Letter 105/2/2014

Applied Statistics II

STA2601

Semester 2

Department of Statistics

TRIAL EXAMINATION PAPER

BAR CODE

Learn without limits.

UNISA
university
of south africa

Dear Student
Congratulations if you obtained examination admission by submitting assignment 1. I would like to take the opportunity of wishing you well in the coming examinations. I hope you found the module stimulating.

The examination

Please note the following with regard to the examination:

- The duration of the examination paper is **two-hours**. You will be able to complete the set paper in 2 hours, but there will be no time for dreaming or sitting on questions you are unsure about. Make sure that you take along a functional scientific calculator that you can operate with ease as it can save you some time. My advice to you would be to do those questions you find easy *first;* then go back to the ones that need more thinking. I do not mind to mark questions in whatever order you do them, just *make sure that you number them clearly!*
- A copy of the list of formulae is attached to the trial examination paper. Please ensure that you know how to test the various hypotheses.
- All the necessary statistical tables will be supplied (see the trial paper).
- Pocket calculators are necessary for doing the calculations.
- Working through (and understanding!) ALL the examples and exercises in the study guide, workbook and in the assignments as well as the trial paper will provide beneficial supplementary preparation.
- Make sure that you know all the theory as well as the practical applications.
- All the chapters in the study guide are equally important and don't try to spot!
- Start preparing early and don't hesitate to call or email me if something is unclear.

Trial paper

Reserve two hours for yourself and do the trial paper under exam conditions on your own!

Duration: 2 hours **100 Marks**

INSTRUCTIONS

1. Answer ALL questions.

2. Marks will not be given for answers only. Show clearly how you solve each problem.

3. For all hypothesis-testing problems always give

   (i) the null and alternative hypothesis to be tested;

   (ii) the test statistic to be used; and

   (iii) the critical region for rejecting the null hypothesis.

4. Justify your answer completely if you make use of JMP output to answer a question.

QUESTION 1

(a) Complete the following:

(i) The statistic T is called an unbiased estimator for the parameter θ if .... (1)

(ii) Let $X_1, ..., X_n$ be a random sample from a population with unknown variance σ^2. An unbiased estimator for the population variance σ^2 is given by $\widehat{\sigma^2} =$ .... (1)

(b) State in general, the three main steps when calculating a maximum likelihood estimator for a parameter θ if the p.d.f. is $f(X;\theta)$. (Give formulae where appropriate.) (4)

[6]

QUESTION 2

Let X_1, X_2, $\ldots$,X_n be a random sample of size n from a discrete distribution with a probability density function

$$P(X = r) = \frac{\lambda^r e^{-\lambda}}{(1 - e^{-\lambda})\, r!} \quad \text{for} \quad r = 1; 2; \cdots$$

[Please note:

In a real life situation a random sample will result in, for example, $X_1 = 2$; $X_2 = 3$; $X_3 = 3$; $X_4 = 10$; $X_5 = 2$ etc$\cdots$.Do not fall into the trap to argue that $X_1 = 1$; $X_2 = 2$; $\cdots$; $X_n = n$ because this is only one very specific outcome out of the millions of other possibilities. Denote the sample outcome by $X_1 = r_1$; $X_2 = r_2$; $\cdots$; $X_n = r_n$.]

(a) Find the likelihood function for the sample. (5)

(b) Show that $\dfrac{\partial InL(\lambda)}{d\lambda} = -n - \dfrac{ne^{-\lambda}}{(1 - e^{-\lambda})} + \dfrac{\sum_{i=1}^{n} r_i}{\lambda}$ (5)

[10]

QUESTION 3

(a) Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with **unknown mean** μ**.** Use the distribution of $U = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{\sigma^2}$ to show that a $100\,(1-\alpha)\,\%$ two-sided confidence interval for σ^2 is given by

$$\left[\frac{\sum \left(X_i - \overline{X}\right)^2}{\chi^2_{\frac{\alpha}{2};n-1}} \quad ; \quad \frac{\sum \left(X_i - \overline{X}\right)^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \right]$$

(5)

(b) Given that X is normally distributed and that a random sample of size $n = 30$ from this distribution yielded the following statistics:

$$\overline{X} = 41 \qquad S = 6$$

(i) Construct a 90% two-sided confidence interval for σ^2. (5)

(ii) How can you use the confidence interval to conclude on a hypothesis test of $\sigma^2 = 30$?(1)

(c) In an investigation of a random sample of 12 epileptic children by E.E.G., the occurrence of cerebral lesions in the anterior and posterior regions of the right and left hemispheres of the brain was noted and the following results were obtained:

		Hemisphere		**Row**
		Right	Left	**total**
Region	Anterior	2	5	7
	Posterior	4	1	5
Column total		6	6	12

Do the data supply evidence of an association between hemisphere and site of lesion? Use $\alpha = 0.05$ (8)

[19]

QUESTION 4

(a) In an experiment to test the effect of heroin on mental activity, 20 voluteers had to complete a questionnaire before given an injection of heroin and again two hours after taking the drug. Their mental activity scores are as follows:

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Score before	7	6	16	12	14	19	8	11	12	18	10	13	11	12	5	15	12	17	9	10
Score after	8	6	15	9	12	16	5	10	7	14	6	14	10	9	4	11	10	15	7	9

The following SAS JMP output was obtained.

Figure 1

Let $Y_i =$ Score after - Score before

(i) Using the output in Figure 1, do the results confirm that heroin *decreases mental activity*? Test at the 1% level of significance and show all your steps. (6)

(ii) Is there any flaw in this experiment? Give a brief description. (2)

(b) A chicken farmer knows from experience that it does not pay him to slaughter his chickens before they weigh more than 2kg on average. He takes a random sample of size $n = 9$ and finds the following slaughtered masses (in kg):

$$1.80 \quad 2.10 \quad 1.90 \quad 2.20 \quad 2.00 \quad 2.10 \quad 1.95 \quad 1.85 \quad 1.80$$

Let μ denote the mean slaughtered mass. Are the chickens ready for slaughtering? The following SAS JMP output was obtained.

Figure 2

Figure 3

You can make use of the SAS JMP output in Figure 2 and Figure 3.

(i) Formulate an appropriate null and alternative hypothesis for this problem. (2)

(ii) What assumption(s) is/are necessary in order to conduct the statistical test?. (2)

(iii) Conduct the appropriate test at the $\alpha = 0.01$ level of significance. State the decision rule and conclusion. (4)

(iii) How will the test procedure change, if in fact you know that $\sigma = 0.10kg$? (4)

[20]

QUESTION 5

A pathologist chooses FOUR different groups for a clinical experiment. Group 1 consists of a random sample of twenty-one (i.e. $n_1 = 21)$ 15-year-old-boys; group 2 consists of a random sample of twenty-one (i.e. $n_2 = 21)$ 15-year-old-girls; group 3 consists of a random sample of twenty-one (i.e. $n_3 = 21)$ males between the ages 35 and 45 years and group 4 consists of a random sample of twenty-one (i.e. $n_4 = 21)$ females between the ages 35 and 45 years. He made sure that there exists no relationship (e.g. parent and child) between a patient of one group and a patient of another group. He measured their cholesterol and found the scores (in mmol/liter) given in the table below:

Cholesterol scores (in mmol/liter)				
	15-year-old boys (X_1)	15-year-old girls (X_2)	Males between 35 and 45 (X_3)	Females between 35 and 45 (X_4)
1	4.26	2.88	4.27	3.06
2	4.20	2.26	4.39	4.53
3	3.55	3.03	5.53	3.46
4	4.24	3.36	5.11	3.77
5	3.70	2.24	4.68	3.55
6	3.13	4.15	4.52	4.85
7	4.01	3.52	4.49	4.69
8	3.52	3.43	5.38	3.86
9	2.68	3.04	4.74	4.09
10	3.56	2.71	4.83	3.64
11	3.88	3.04	4.85	4.33
12	2.66	2.68	4.86	4.92
13	3.15	4.32	4.46	4.85
14	4.03	2.58	3.90	4.36
15	4.37	3.15	3.66	4.90
16	4.10	2.40	4.37	4.61
17	3.12	4.02	3.36	3.00
18	3.69	3.26	5.16	4.18
19	3.48	3.35	3.99	4.90
20	2.81	3.21	5.01	3.55
21	3.46	2.47	5.04	5.10

You may MAKE USE of the following calculations:

$$\begin{array}{llllll}
\sum_{j=1}^{21} x_{1j} = 75.60 & & \sum_{j=1}^{21} (X_{1j})^2 = 277.7100 & & \sum_{j=1}^{21} (x_{1j} - \overline{x}_1)^2 = 5.5500 \\
\sum_{j=1}^{21} x_{2j} = 65.10 & & \sum_{j=1}^{21} (X_{2j})^2 = 208.6944 & & \sum_{j=1}^{21} (x_{2j} - \overline{x}_2)^2 = 6.8844 \\
\sum_{j=1}^{21} x_{3j} = 96.60 & & \sum_{j=1}^{21} (X_{3j})^2 = 450.5050 & & \sum_{j=1}^{21} (x_{3j} - \overline{x}_3)^2 = 6.1450 \\
\sum_{j=1}^{21} x_{4j} = 88.20 & & \sum_{j=1}^{21} (X_{4j})^2 = 378.9118 & & \sum_{j=1}^{21} (x_{4j} - \overline{x}_4)^2 = 8.4718
\end{array}$$

(a) Compute **an unbiased estimate** of the population variance of each clinical group (4)

(b) (i) Compute the “ordinary” average of the four variances computed in (a) (2)

(ii) Compute the MSE according to the definition in the study guide. What do you notice?(4)

(c) Do you think it is reasonable to assume that the assumption of independence is met? Substantiate. (2)

(d) Test at the 5% level of significance whether there is a difference in the mean cholesterol scores of the four different groups.

(i) State the null and alternative hypotheses.

(ii) State the rejection region and conclusion.

(11)

[23]

QUESTION 6

The maintenance cost of computer equipment seems to increase with the age of the equipment. The following data were collected from a random sample:

	Age (years)	**Cost** (R)			
	x_i	y_i	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
	0.75	110	2.7225	21 316	240.90
	1.00	150	1.9600	11 236	148.40
	1.25	130	1.3225	15 876	144.90
	1.75	205	0.4225	2 601	33.15
	2.25	235	0.0225	441	3.15
	2.75	290	0.1225	1 156	11.90
	3.00	330	0.3600	5 476	44.40
	3.25	310	0.7225	2 916	45.90
	3.75	385	1.8225	16 641	174.15
	4.25	415	3.4225	25 281	294.15
Total	**24.00**	**2 560**	**12.9000**	**102 940**	**1 141.00**

(a) Compute the sample correlation coefficient, r. (3)

(b) Test $H_0 : \rho = 0.8$ against $H_1 : \rho > 0.8$ at the 1% level of significance. (6)

(c) Is this an ideal (causal) regression experiment? Comment. (2)

(d) Assume that a linear relationship $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where the ε_i's are independent $n(0; \sigma^2)$ random variables, is meaningful. Estimate $\beta_0, \ \beta_1$ and hence give the regression line of y on x. (4)

(e) What is the predicted cost for equipment that is 5 years old? (1)

(d) Find a 95% confidence interval for the predicted cost for equipment that is 5 years old. Given that $S^2 = 252.37375.$ (6)

[22]

[100]

Formulae / Formules

$$B_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2]^{\frac{3}{2}}}$$

$$B_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2]^2}$$

$$A = \frac{\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

$$\rho = \frac{e^{\eta} - e^{-\eta}}{e^{\eta} + e^{-\eta}}$$

$$T = \sqrt{n-2}\frac{U_{11} - U_{22}}{2\sqrt{U_{11}U_{22} - U_{12}^2}}$$

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\upsilon = \frac{\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right]^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}$$

$$F = \frac{n\sum_{i=1}^{k}(\overline{X}_i - \overline{X})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2/(kn-k)}$$

$\widehat{\beta}_1 = \dfrac{\sum_{i=1}^{n} Y_i(X_i - \overline{X})}{d^2}$ Note: $d^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$ and $\widehat{\beta}_0 = \dfrac{\sum_{i=1}^{n} Y_i - \widehat{\beta}_1 \sum_{i=1}^{n} X_i}{n} = \overline{Y} - \widehat{\beta}_1\overline{X}$

Table A. Percentage points for the distribution of B_1
Lower percentage point $=$ – (tabulated upper percentage point)

Size of sample	Percentage points	Size of sample	Percentage points
n	5%	n	5%
25	$0,711$	200	$0,280$
30	$0,662$	250	$0,251$
35	$0,621$	300	$0,230$
40	$0,587$	350	$0,213$
45	$0,558$	400	$0,200$
50	$0,534$	450	$0,188$
		500	$0,179$
60	$0,492$	550	$0,171$
70	$0,459$	600	$0,163$
80	$0,432$	650	$0,157$
90	$0,409$	700	$0,151$
100	$0,389$	750	$0,146$
		800	$0,142$
125	$0,350$	850	$0,138$
150	$0,321$	900	$0,134$
175	$0,298$	950	$0,130$
200	$0,280$	1000	$0,127$

Table B. Percentage points of the distribution of B_2

Size of sample n	Percentage points	
	Upper 5%	Lower 5%
50	$3,99$	$2,15$
75	$3,87$	$2,27$
100	$3,77$	$2,35$
125	$3,71$	$2,40$
150	$3,65$	$2,45$
200	$3,57$	$2,51$
250	$3,52$	$2,55$
300	$3,47$	$2,59$
350	$3,44$	$2,62$
400	$3,41$	$2,64$
450	$3,39$	$2,66$
500	$3,37$	$2,67$
550	$3,35$	$2,69$
600	$3,34$	$2,70$
650	$3,33$	$2,71$
700	$3,31$	$2,72$
800	$3,29$	$2,74$
900	$3,28$	$2,75$
1000	$3,26$	$2,76$

Table C. Percentage points for the distribution of $A = \dfrac{\text{mean deviation}}{\text{standard deviation}}$

Size of sample n	$n-1$	Percentage points			
		Upper 5%	Upper 10%	Lower 10%	Lower 5%
11	10	0,9073	0,8899	0,7409	0,7153
16	15	0,8884	0,8733	0,7452	0,7236
21	20	0,8768	0,8631	0,7495	0,7304
26	25	0,8686	0,8570	0,7530	0,7360
31	30	0,8625	0,8511	0,7559	0,7404
36	35	0,8578	0,8468	0,7583	0,7440
41	40	0,8540	0,8436	0,7604	0,7470
46	45	0,8508	0,8409	0,7621	0,7496
51	50	0,8481	0,8385	0,7636	0,7518
61	60	0,8434	0,8349	0,7662	0,7554
71	70	0,8403	0,8321	0,7683	0,7583
81	80	0,8376	0,8298	0,7700	0,7607
91	90	0,8353	0,8279	0,7714	0,7626
101	100	0,8344	0,8264	0,7726	0,7644

Table D
Tabel D

The hypergeometric probability distribution: $P\left(X \leq x\right)$ for $N = 12$

Die hipergeometriese verdeling: $P\left(X \leq x\right)$ vir $N = 12$

n	k	x	P	n	k	x	P	n	k	x	P
1	1	0	0,917	4	4	0	0,141	6	2	0	0,227
1	1	1	1,000	4	4	1	0,594	6	2	1	0,773
				4	4	2	0,933	6	2	2	1,000
2	1	0	0,833	4	4	3	0,998				
2	1	1	1,000	4	4	4	1,000	6	3	0	0,091
								6	3	1	0,500
2	2	0	0,682	5	1	0	0,583	6	3	2	0,909
2	2	1	0,985	5	1	1	1,000	6	3	3	1,000
2	2	2	1,000								
				5	2	0	0,318	6	4	0	0,030
3	1	0	0,750	5	2	1	0,848	6	4	1	0,273
3	1	1	1,000	5	2	2	1,000	6	4	2	0,727
								6	4	3	0,970
3	2	0	0,545	5	3	0	0,159	6	4	4	1,000
3	2	1	0,955	5	3	1	0,636				
3	2	2	1,000	5	3	2	0,955	6	5	0	0,008
				5	3	3	1,000	6	5	1	0,121
3	3	0	0,382					6	5	2	0,500
3	3	1	0,873	5	4	0	0,071	6	5	3	0,879
3	3	2	0,995	5	4	1	0,424	6	5	4	0,992
3	3	3	1,000	5	4	2	0,848	6	5	5	1,000
				5	4	3	0,990				
4	1	0	0,667	5	4	4	1,000	6	6	0	0,001
4	1	1	1,000					6	6	1	0,040
				5	5	0	0,027	6	6	2	0,284
4	2	0	0,424	5	5	1	0,247	6	6	3	0,716
4	2	1	0,909	5	5	2	0,689	6	6	4	0,960
4	2	2	1,000	5	5	3	0,955	6	6	5	0,999
				5	5	4	0,999	6	6	6	1,000
4	3	0	0,255	5	5	5	1,000				
4	3	1	0,764								
4	3	2	0,982	6	1	0	0,500				
4	3	3	1,000	6	1	1	1,000				

Table E

Upper 5% percentage points of the ratio, $S^2_{\max}/S^2_{\min}$

v	$k = 2$	3	4	5	6
2	$39,0$	$87,5$	142	202	266
3	$15,4$	$27,8$	$39,2$	$50,7$	$62,0$
4	$9,60$	$15,5$	$20,6$	$25,2$	$29,5$
5	$7,15$	$10,8$	$13,7$	$16,3$	$18,7$
6	$5,82$	$8,38$	$10,4$	$12,1$	$13,7$
7	$4,99$	$6,94$	$8,44$	$9,70$	$10,8$
8	$4,43$	$6,00$	$7,18$	$8,12$	$9,03$
9	$4,03$	$5,34$	$6,31$	$7,11$	$7,80$
10	$3,72$	$4,85$	$5,67$	$6,34$	$6,92$
12	$3,28$	$4,16$	$4,79$	$5,30$	$5,72$
15	$2,86$	$3,54$	$4,01$	$4,37$	$4,68$
20	$2,46$	$2,95$	$3,29$	$3,54$	$3,76$
30	$2,07$	$2,40$	$2,61$	$2,78$	$2,91$
60	$1,67$	$1,85$	$1,96$	$2,04$	$2,11$
∞	$1,00$	$1,00$	$1,00$	$1,00$	$1,00$

$k =$ number of samples

$v =$ degrees of freedom for each sample variance