NATIONAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

FACULTY OF INDUSTRIAL TECHNOLOGY

DEPARTMENT OF TECHNICAL TEACHER EDUCATION

Programme: BACHELOR OF TECHNICAL EDUCATION HONOURS DEGREE

GENERAL EXAMINATION

Course: EDUCATIONAL STATISTICS **TTE3009**

Part/Year: III **MAY 2006**

Time: 3 hours **100 marks**

DIRECTIONS TO CANDIDATES

1. Answer **Question 1** and any **THREE** others.
2. All questions carry equal marks.
3. Each question should begin on a fresh page and parts of the same question must be together.
4. Formulae are provided at the back of the question paper.
5. Use of silent scientific calculators is encouraged. Indicate by writing “By Calculator…”
6. Statistical tables are supplied in a separate booklet.

QUESTION 1

Ten joints of some foodstuff are each cut in half; one half is frozen and wrapped by process A and the other half by a new process B. The halves are placed in ten freezers, halves of the same joint being put in the same freezer. The number of days to spoilage, which can be detected from a change in the colour of a pack, are found for each pack as indicated in the table below:

Joint number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Process A	63	109	82	156	161	155	47	141	92	149
Process B	129	105	76	207	253	146	62	160	90	177

(a) Calculate the means for processes A and B. [4]

(b) Assuming that the data above represent a random sample from a normal distribution, use a 5% significance level to test the hypothesis that the wrapping processes A and B are equally effective in preserving the freshness of the meat. [10]

(c) State any three assumptions that you make in this experiment. [3]

(d) Explain clearly a method of obtaining a random sample. Use examples. [8]

QUESTION 2

(a) The table below gives the marks obtained by a class in a physics examination.

Marks	49 – 51	52 – 54	55 – 57	58 – 60	61 – 63	64 – 66
Frequency	4	6	12	10	6	2

(i) Draw a cumulative frequency curve from the data.
(ii) From the curve, indicate the median mark.
(iii) Estimate also the number of students passing if the pass mark was set at 55.
(iv) Indicate on the graph and state the upper and lower quartiles.
(v) Describe the *semi-interquartile range* and give its value in this distribution. [21]

(b) Compare *absolute* and *relative* frequency. [4]

QUESTION 3

For the set of data below,

X	7	10	8	5	11	3	7	11	12	6
Y	2.0	3.0	2.4	1.8	3.2	1.5	2.1	3.8	4.0	2.2

a) Plot the scatter diagram. [6]
b) Find and insert the best fitting line. [2]
c) Predict Y for X = 12, 14, and 18 using the line. [3]
d) Compute both Pearson's product-moment correlation coefficient and Spearman's rank correlation coefficient. [10]
e) Comment fully on the meaning of the coefficients obtained. [4]

QUESTION 4

The amounts, in millions of dollars, of expenditure on services in a city for the years 1991 and 1992 are shown in the table below.

	Expenditure ($million)	
	1991	**1992**
Education	150	x
Emergency services	121.5	y
Roads	83	98.5
Social services	72.5	92
Other services	59	79.5
TOTAL	486	z

Pie charts are drawn to compare the expenditure for these two years.

a) Find the angle representing the Education sector in 1991. [5]
b) Given that there is a 20% increase in expenditure on education for the year 1992 and that the angles of the sectors representing Emergency Services are the same for the two years, calculate the values of x, y and z. [5]
c) Given also that the radius of the pie chart for 1991 is 9 cm, calculate the radius of the pie chart for 1992. [5]
d) Draw the two pie charts to scale on plain paper. [10]

QUESTION 5

(a) The best times, to the nearest second, that the 50 members of an athletics club have achieved in running 1500 m are grouped in the table below.

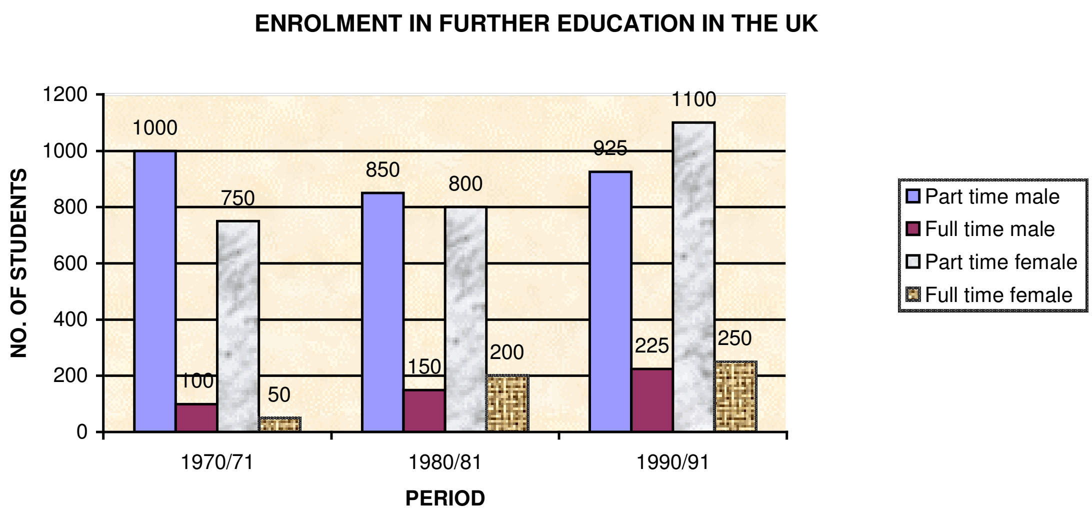
Time (seconds)	240 – 249	250 –259	260 – 264	265 – 274	275 - 294
No. of athletes	8	14	9	11	8

(i) Draw an appropriate histogram to illustrate these data.
(ii) State the modal class centre.
(iii) Calculate the mode for the data. [10]

(b) Explain the Pareto principle (or the 80/20 rule), giving an example. [5]
(c) Define and compare the arithmetic and the geometric means. [4]
(d) Explain the advantages and disadvantages of using the mean, median and mode as measures of central tendency. [6]

QUESTION 6

(a) (i) In the graph below, state all the variables shown, and say whether they are dependent or independent, continuous or discrete, qualitative or quantitative. [6]



(ii) Calculate the percentage change in the proportion of enrolled females in 1990/91 compared to 1980/81. [6]

(b) In a certain country the heights of adult males have a mean of 170cm and a standard deviation of 10cm, and the heights of adult females have a mean of 160cm and a standard deviation of 8cm; for each gender group the distribution of heights approximates closely to a normal probability model. On the hypothesis that height is not a factor in selecting a marriage partner, calculate the probability that a husband and wife selected at random are both taller than 164cm. [13]

END OF EXAMINATION

LIST OF SELECTED FORMULAE

Sample mean $\bar{x} = \frac{\sum fx}{n}$ **Weighted mean** $\bar{x}_w = \frac{\sum wx}{\sum w}$

Geometric Mean = $\sqrt[n]{x_1 x_2 x_3 \cdots\cdots x_n}$

Sample Standard deviation: For ungrouped data **s** **=** $\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

: For grouped data **s** **=** $\sqrt{\frac{\sum f(x-\bar{x})^2}{n-1}}$

Binomial distribution : P(r) $= {}^nC_r p^r q^{n-r}$
Mean (μ) = np
Standard deviation = $\sqrt{npq}$

Pearson's product-moment correlation coefficient (r)

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

Spearman's rank correlation coefficient (ρ, *rho* or r_s)

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where d = differences in ranks of paired scores

Sample Coefficient of determination (r^2)

$$r^2 = \frac{a\sum Y + b\sum XY - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}$$

Where: a = Y- intercept
b = gradient of regression line
X = value of independent variable
Y = value of dependent variable
$\bar{Y}$ = mean of Y values

Chi-square (χ^2) test

$$\chi^2 = \sum \frac{(F_o - F_e)^2}{F_e}$$

Where F_o= observed frequency
F_e= expected frequency

For contingency tables in each cell, $F_e = \frac{RowTotal \times ColumnTotal}{n}$

z-score (for a normal distribution N(μ, σ^2)

$$z = \frac{x-\mu}{s}$$

Where x = score
μ = mean
s = standard deviation
z = number of standard deviation units from x to μ

The t-test

(a) For two groups:

$$t_{(n_1 + n_2 - 2)} = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}}$$

Where $\overline{X}_1$ and $\overline{X}_2$ = means for group 1 and 2
s = standard deviations
n = number in group

(b) For comparing a sample and a population:

$$t_{(n-1)} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

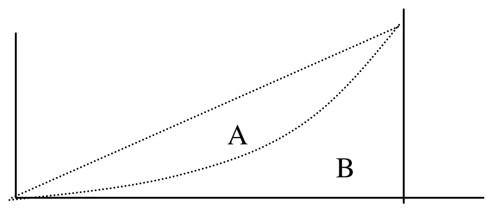
Where $t_{(n-1)}$ = t value for n - 1 degrees of freedom
$\overline{X}$ = sample mean
μ = population (or expected) mean
s = sample standard deviation

(c) For paired comparisons (e.g. pretest-posttest)

$$t_{(n-1)} = \frac{\overline{D} - \mu_0}{\frac{s_D}{\sqrt{n}}}$$

where $\overline{D}$ = mean difference between groups
μ_0 = expected mean difference (usually = 0)
s_D = standard deviation of differences
n = number of members/items in sample

The Lorenz Curve



Gini coefficient

$$= \frac{area\,below\,line\,of\ equal\ distribution - area\,below\,curve}{area\,below\,curve}$$

$$= \frac{area\,A}{area\,B}$$

Percentile Point (P_p)

$$P_p = LL_i + \left[\left(\frac{n_p - C_f}{f_i}\right) \times I\right]$$

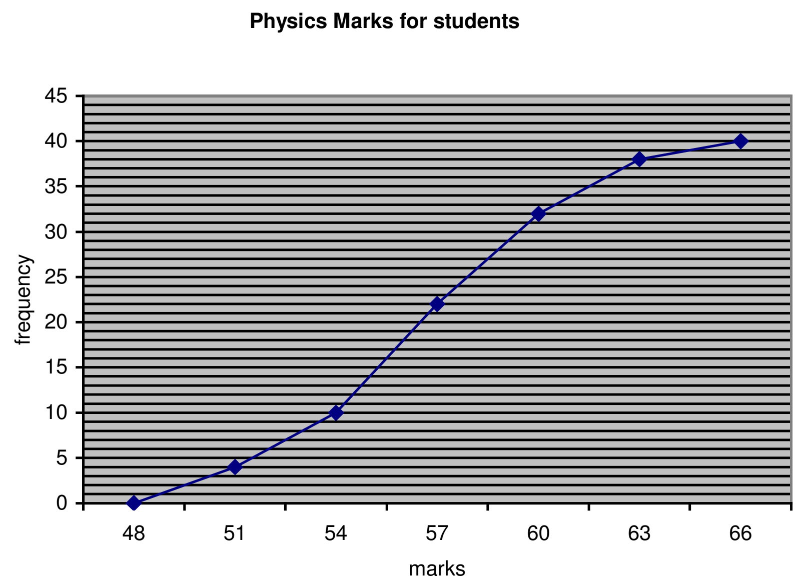
where P_p = the p^th^ percentile point (eg 65^th^)
LL_i = exact lower limit of percentile interval
n_p = number of cases in specified % of N
C_f = cumulative frequency up to percentile interval
f_i = frequency within the percentile
I = class interval

MARKING GUIDE : TTE3009 EDUCATIONAL STATISTICS: MAY 2006

Question	Marking Points	Marks
1 (a)	Process A: mean = $\frac{\sum x_1}{n} = \frac{1155}{10}$ = **<u>115.5</u>** [2] Process B: mean $= \frac{1405}{10}$ = **<u>140.5</u>** [2]	**4**
1 (b)	Process A (s.d. = 57.5 by calculator) and Process B (s.d. = 40.26) [4] t-test $t_8 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}} = \frac{115.5 - 140.5}{\sqrt{\frac{57.5}{10} + \frac{40.26}{10}}} = \frac{25}{\sqrt{9.776}}$ **= 8.000** [4] Calculated t (8.000) > t (1.73) at 0,05 sig. level and 18 d.f. in the tables Therefore there is a significant difference in the performance of the two processes. [2]	**10**
1 (c)	Assumptions - Food halves are identical in composition/mass/etc - food halves positioned the same in the freezer - uniform conditions in the freezer, etc [3]	**3**
1 (d)	Random sampling methods (full explanation with examples) – e.g. raffle/lottery/coin toss/ random numbers etc	**8**

Question	Marking Points	Marks
2 (a)	(i) [3]	21

Marks	46 - 48	49 – 51	52 – 54	55 – 57	58 – 60	61 – 63	64 – 66
Cutoff points	48	51	54	57	60	63	66
Frequency	0	4	6	12	10	6	2
Cum Freq	0	4	10	22	32	38	40



[6]

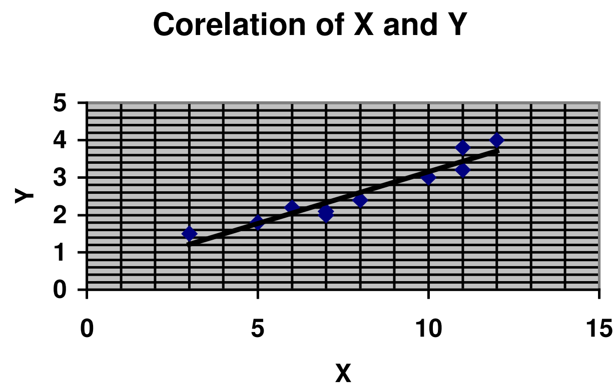
(ii) The middle of the range is between student 20 and student 21 i.e. 20.5
From the graph the mark corresponding to 20,5 is 56.5
∴ **Median = 56.5 (or 55.5)** [3]

(iii) When the mark is 55, the cumulative frequency is 13. So 13 students scored 55 or less, and therefore 12 students scored less than 55, and the rest passed.
∴Number of students who passed = 40 – 12 = 28 (or 23) [3]

(iv) The lower quartile is between students 10 and 11, i.e. 10.5
The upper quartile is between students 30 and 31 i.e. 30.5
From the graph, Lower quartile Q_1 = **54 (or 53)**
Upper quartile Q_3 = **59.5 (or58.5)** [3]

(v) The semi-interquartile range is : also called quartile deviation
: the value halfway between Q_1 and Q_3
: formula is ½ $(Q_3 - Q_1)$
= **½(5.5)**
= **2.75** [3]

Question	Marking Points	Marks
2 (b)	b) **Absolute frequency** is the actual number of cases per class e.g. 8 students in a group of 40 pass with distinctions. **Relative frequency** is the number of cases per class as a percentage or proportion of the whole group e.g. 20% (or 0.2) females in a group.	4

Question	Marking Points	Marks
3 (a) & (b)		6 2
3 (c)	When X = 12, Y = **3.65** (approx) X = 14, Y = **4.22** X = 18, Y = **5.31**	3
3 (d)	(see below)	10

3 (d)

X	Y	X^2	Y^2	XY	R_X	Or R_X	R_Y	Or R_Y	$\|d\|$	d^2
7	2.0	49	4.00	14.0	6.5	4.5	8	3	1.5	2.25
10	3.0	100	9.00	30.0	4	7	4	7	0	0
8	2.4	64	5.76	19.2	5	6	5	6	0	0
5	1.8	25	3.24	9.0	9	2	9	2	0	0
11	3.2	121	10.24	35.2	2.5	8.5	3	8	0.5	0.25
3	1.5	9	2.25	4.5	10	1	10	1	0	0
7	2.1	49	4.41	14.7	6.5	4.5	7	4	0.5	0.25
11	3.8	121	14.44	41.8	2.5	8.5	2	9	0.5	0.25
12	4.0	144	16.00	48.0	1	10	1	10	0	0
6	2.2	36	4.84	13.2	8	3	6	5	2	4.00
$\sum X = 80$	$\sum Y = 26$	$\sum X^2 = 718$	$\sum Y^2 = 74.18$	$\sum XY = 229.6$						$\sum d^2 = 7.00$

[4]

Pearson's (p-m) correlation **r** =

$$\frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

= **<u>0.9534</u>** (by calculator) [3]

Spearman's rank correlation $\mathbf{r_s}$ = $1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6.7}{10.99}$

= **<u>0.9576</u>** (by calculator) [3]

Question	Marking Points	Marks
3 (e)	Both coefficients are positive and high, showing that X and Y are closely correlated, and one variable can be used reliably to predict the other. An increase in X is accompanied by an increase in Y.	4

Question	Marking Points	Marks

4 (a)	$150/486 \times 360^0 =$ <u>**111⁰**</u>	**5**

4 (b)

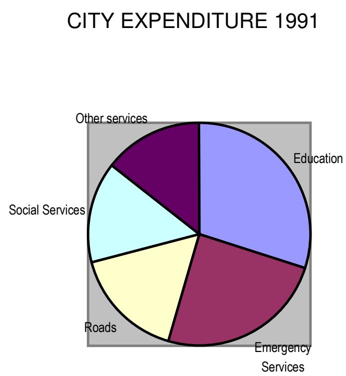
	Expenditure ($m)		Angles(^{0})	
	1991	**1992**	**1991**	**1992**
Education	150	x = 180	111.1	108
Emergency services	121.5	y = 150	90	90
Roads	83	98.5	61.5	59.1
Social services	72.5	92	53.7	55.24
Other services	59	79.5	43.7	47.7
TOTAL	486	z =600	360	360
TOTAL EXPENDITURE	486	z =600		
Area ratio	1	1.23		
Radius ratio	1	1.1		
Radius (cm)	9	10		

x = $180 000 000 **y = $150 000 000** **z = $600 000 000**

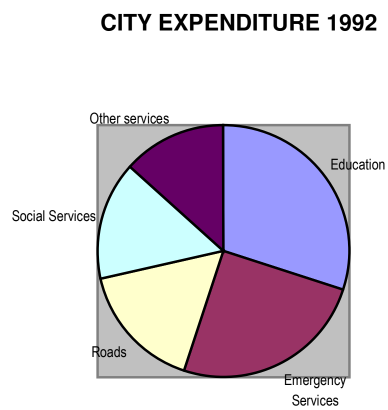
5

4 (c)	Radius for 1992 = **10 cm** (Use calculations in bottom shaded area of table in 4(b)	**5**

4 (d) For calculation of angles see last two shaded columns of 4(b).

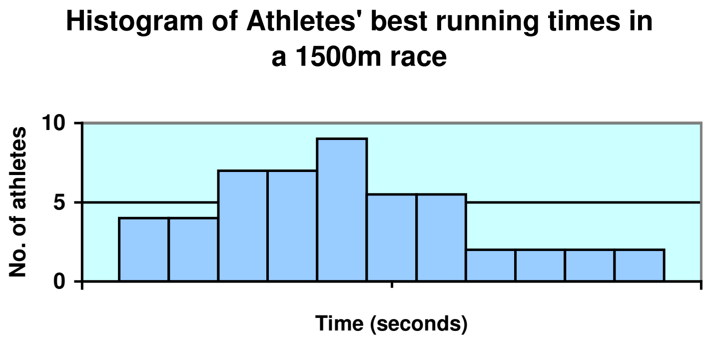


Radius = 9 cm

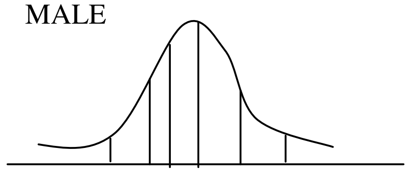
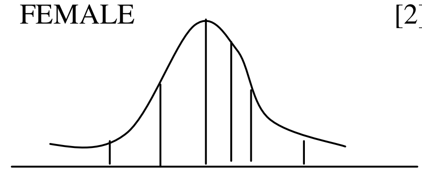


Radius = 10 cm [[10]

10

Question	Marking Points	Marks
5 (a)	(i)  [5] (i) Modal class centre: =**262** (Explanation: frequency of 9 is higher in a class interval of 5 seconds, compared to other wider class intervals) [1] (ii) Mode (for grouped data) $M_o = 259.5 + \frac{-5}{-5+-2}5$ $= 259.5 + 3.57$ $=$ **263.07** [4]	**10**
5 (b)	(b) Pareto principle - when 20 % of one variable accounts for 80% of the other - e.g. 20% of people own 80% of wealth - represented in a Lorenz curve - used to demonstrate distribution pattern of resources	**5**
5 (c)	Arithmetic mean - the average of any variable/ the sum of values divided by their number Geometric mean – average of the rate of increase/change, e.g. year on year	**4**
5 (d)	Mean – Adv - calculated using all the values - more stable for further calculations - Disadv- may be inaccurate in a skewed distribution Median - Adv - useful for large no. of objects - Disadv- unreliable Mode - Adv - easy to locate/estimate - Disadv - unreliable, may be multimodal	**6**

<table>
<tr><th>Question</th><th>Marking Points</th><th>Marks</th></tr>
<tr><td>6 (a)</td><td>(i) Period(years) - independent – discrete – qualitative
Gender - independent – discrete – qualitative
Mode of study - independent - discrete - qualitative
Frequency - dependent - discrete - quantitative [½ mark each] [6]

(ii) Proportion of enrolled females (1980/81) = $\dfrac{800+200}{800+200+850+150} = \dfrac{1000}{2000} = 0.5$
Proportion of enrolled females (1990/91) = $\dfrac{1100+250}{1100+250+925+225} = \dfrac{1350}{2500} = 0.54$
$\therefore$ % change = $\dfrac{0.54-0.50}{0.50}$ x 100 = <u>8% increase</u>
[6]</td><td>12</td></tr>
<tr><td>6 (b)</td><td>MALE

FEMALE [2]


z-score for males 164cm tall = $\dfrac{164-170}{10}$ = - 0.6
$\therefore$ probability of males taller than 164cm = 1 - 0.2743 (from tables)
= <u>0.7257</u> [4]

z-score for females 164cm tall = $\dfrac{164-160}{8}$ = 0.5
$\therefore$ probability of females taller than 164cm = 1 - 0.6915 (from tables)
= <u>0.3085</u> [4]

$\therefore$ Probability of a randomly selected couple both taller than 164cm
= 0.7257 x 0.3085
<u>= 0.2239</u> [3]</td><td>13</td></tr>
</table>