SES 865: Environmental Statistics
MAIN EXAMINATION

MAASAI MARA UNIVERSITY
School of Tourism & Natural Resource Management
Department of Environment, Forestry & Agriculture

SECTION A: Answer all questions

Qn 1: What is a variable? Differentiate between quantitative and qualitative data. (3mks)

Qn 2: A Census Bureau report gave data on the number of times residents of Narok County had been married, for subjects of various ages. Table below summarizes responses for subjects of age 20 to 24. The frequencies are in thousands of people, for instance, 7,074,000 men never married.

Table 1: Number of Times Married, for Subjects of Age 20 to 24.

Number of Times Married	Women	Men
0	5861	7074
1	2773	1541
2	105	43

a) Find the median and the mean for each gender. Explain your answer. Which of these two measures of location is particularly more informative (5mks)

b) Compute the standard deviation and the interquartile range for each gender. Which of these two measures of dispersion is particularly more informative (5mks)

c) The following migration distances corresponding to distance moved by Pastoralist in search of pasture in kilometers were recorded as follows 43, 6, 7, 11, 122, 41, 21, 17, 1, 3. Find the mean and standard deviation and then convert the first 3 observations to ***Z***-score. (7mks)

Qn 3: Given a simple regression with slope b = 3, S_y= 8, and S_x= 2, find the standard error of the estimate (5mks)

SECTION B: Answer any 3 questions

Qn 4: A survey of the Native and immigrants population in Narok County reveals the following annual trip frequencies to the Mara National park.

$$\bar{x}_1 = 4.1, \quad S_1^2 = 14.3, \quad n_1 = 20$$

$$\bar{x}_2 = 3.1, \quad S_2^2 = 12, \quad n_2 = 16$$

a) Assume that the variances are equal and test the null hypothesis that there is no difference between the park going frequencies of Native and immigrant population (5mks)
b) Repeat the exercise assuming that the variances are unequal (5mks)

c) Find the p-value associated with the test in part (a) and (b) (2mks)
d) Find a 95% confidence interval for the difference in means (3mks)

Qn 5: In an experiment to generate 500 data points, two simulations were done. The simulated variables were named Y1 and Y2 respectively. Please answer the following questions sequentially:

(a) Table below provide numerical summary of the two data sets. How useful are these summaries in describing the distributions for Y1 and Y2, respectively? Under what additional assumptions, do these summaries become useful? (4mks)

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Y1	**-4.804**	**3.369**	**6.198**	**6.133**	**9.129**	**17.320**
Y2	**-5.3900**	**0.5533**	**8.8680**	**6.4840**	**12.0400**	**17.3900**

(b) Figure 1 below shows side-by-side notched boxplots for Y1 and Y2. Do these notched boxplots provide more information of these variables than the numerical summaries obtained in Part (a)? Are you totally confident with the graphical summary of the data based on the notched boxplots? Provide a detailed discussion. (4mks)

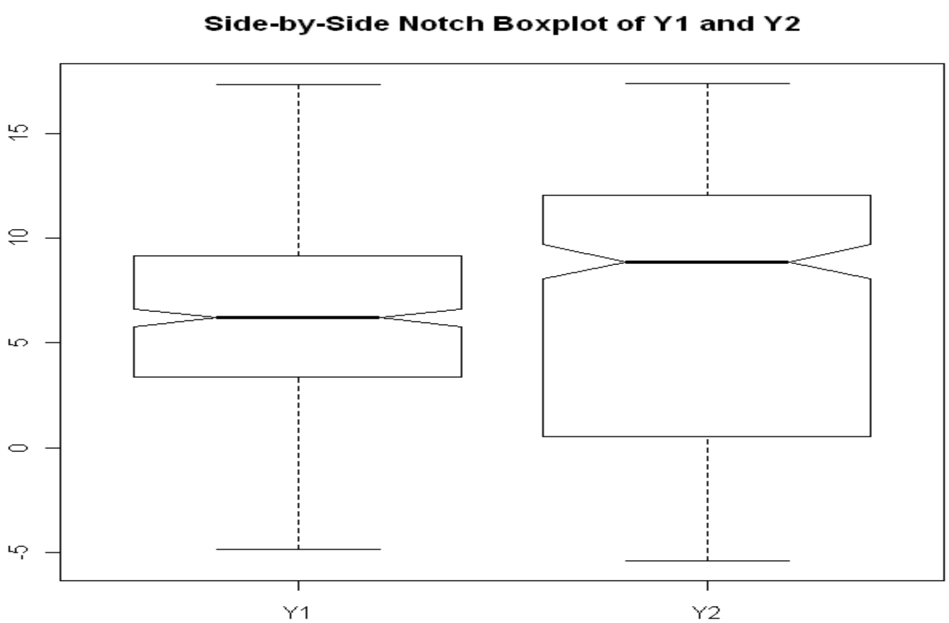

Figure 1: side by side notched box plot of simulation data for variables Y1 and Y2.

(c) Figure 2 below shows quantile plots for Y1 and Y2. Do the quantile plots provide more information of these variables than the notched boxplots? Are there any noticeable differences between these two data distributions? (4mks)

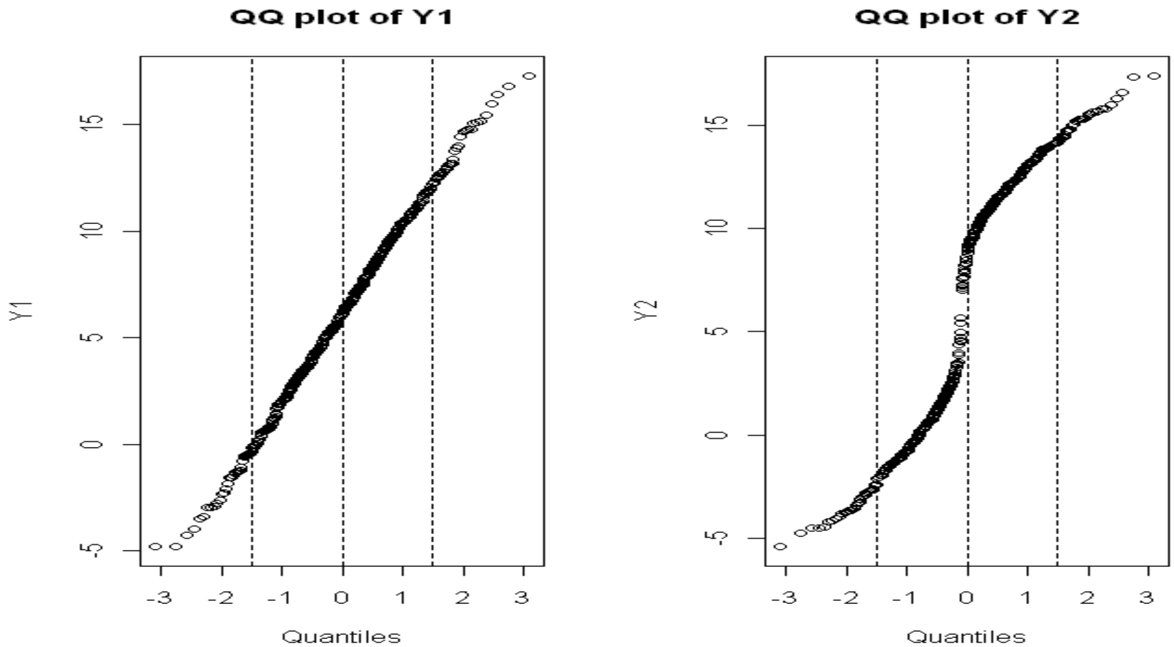

Figure 2: quantile plot of Y_1 and Y_2 from simulated data.

(d) Provide a detailed discussion on the power of each technique used in part (a) to (c). What suggestions would you like to make in order to numerically or graphically summarize the data for a real problem in which no one knows the true distribution of the data? (3mks)

Qn 6: Using the following data,

Table 2: Annual swimming frequencies for three Kenyan cities

	Annual swimming frequencies		
	Nairobi	Kisumu	Mombasa
	38	58	80
	42	66	70
	50	80	60
	57	62	55
	80	73	72
	70	39	73
	32	73	81
	20	68	50
Mean	**48.63**	**63.63**	**67.63**
Standard deviation	**19.88**	**12.66**	**11.43**
X_{++} = 59.96; S^2 = 16.69			

a) Perform an analysis of variance to test that the swimming frequencies does no vary between the three cities. Show the between and within sum of squares, the observed F-statistic and the critical F value (10mks)

b) Use Levene's test to determine whether the assumption of homoscedasty is justified (5mks)

Qn 7: (a) find the correlation coefficient r, for the following sample data on education and income (2mks)

Observation	Income (Ksh)	Education (Years)
1	30	12
2	28	12
3	52	18
4	40	16
5	35	16

(b). Test the null hypothesis that $\rho = 0$ (5mk)

(c). Find the spearman rank correlation for these data (5mks)

(d) Test whether the observed value of $\boldsymbol{r_s}$ is significantly different from zero (3mks)