UNIVERSITY OF SWAZILAND

FINAL EXAMINATION PAPER

PROGRAMME: DIPLOMA IN AGRICULTURE II
DIPLOMA IN AGRICULTURAL EDUCATION II
DIPLOMA IN HOME ECONOMICS II
DIPLOMA IN HOME ECONOMICS EDUCATION II
REMEDIAL IN AGRICULTURE

COURSE CODE: AEM 201

TITLE OF PAPER: ELEMENTARY STATISTICS

TIME ALLOWED: TWO (2) HOURS

INSTRUCTION:
1. QUESTION 1 IS COMPULSORY
2. ANSWER ANY TWO OTHER QUESTIONS
   FOR QUESTION 3 (attempt either 3.1 or 3.2)
   FOR QUESTION 4 (attempt either 4.1 or 4.2)
3. SHOW ALL WORKINGS
4. ONLY INDICATED FORMULAE AUTHORISED

DO NOT OPEN THIS PAPER UNTIL PERMISSION HAS BEEN GRANTED BY THE CHIEF INVIGILATOR

Question 1 (overall=40 marks)

1.1 (10 marks)

Suppose that, in a data collection operation, the following variables have been recorded for several individual people: (a) age, (b) year of birth, (c) sex, (d) height, (e) weight, (f) method of irrigation, (g) , (h) surname, (j) age rank in household (i.e. whether they are oldest, second oldest, …., youngest person in the household where they live), (k) whether they hold a land tenure title. For each of the variables (a) – (k), state whether it is qualitative or quantitative; and if quantitative, state whether it is discrete or continuous.

1. 2 (1.2. a = 5 marks : 1.2.b = 10 marks: 1.2.c=5 marks: 1.2.d=10 marks)

In the following table x is the number of grams of impurity in one litre containers of a chemical solution

x	0-25	26-50	51-75	76-100	101-125	126-150	151-175	176-200
f	20	73	85	114	106	54	36	12

1.2.a) Draw a frequency polygone

1.2.b) Estimate the mean, mode and median impurity content

1.2.c) Drawing from the estimates found in 1.2.b, what indicator would you suggest to summarise the degree of impurity?

1.2.d) Assess the extent to which the distribution is skewed by computing the appropriate indicators.

Question 2 (overall = 25 mark)

2.1 (overall: 15 marks 2.1.a=1 marks: 2.1.b=12 marks: 2.1.c=2 marks)

Data were collected on the number of tractors per farming unit in a certain African country at two different dates in order to assess temporal changes, 1991 and 2001

Number of tractors	Percentage	
	1991	2001
1	1	2
2	5	4
3	12	9
4	27	23
5	35	30
6	13	23
7	3	5
8	4	4

(*Source*: Office of Population Censuses and Surveys)

(2.1.a) Indicate the main variable of study for this data

(2.1.b) For each distribution calculate the mean, standard deviation and coefficient of variation of the number of tractors per farming unit.

(2.1.c) Comment on your results with special reference to the standard deviation and the coefficient of variation.

2.2 (overall =10 marks: 2.2.a = 8 marks : 2.2.b=2 marks)

The following data gives the weight of 1200 duck eggs

Weight (mid-point in grams)	No.of eggs
57	7
60	13
63	68
66	144
69	197
72	204
75	208
78	160
81	101
84	54
87	25
90	13
93	4
96	2

a) Find the median, the first and the third quartile using an interpolation formula method.

b) Comment on your results in terms of percentage concentration.

Question 3 (Attempt either 3.1 = 25 marks or 3.2 = 25 marks)

3.1 (25 marks)

The following data gives the maize yield (x) in millions of tons against the area planted (y) in millions of acres for Southern Africa for the successive years 1984 to 1992.

x	3.7	4.1	3.4	3.8	3.4	3.3	4.2	4.7	4.7
y	2.2	2.5	2.2	2.3	2.4	2.1	2.5	2.7	2.8

(a) Find the regression line of X on Y.

(b) Find the correlation coefficient between X and Y.

(c) Comment on your results.

3.2 (25 marks)

In a nutritional experiment, a number of cultures were subjected to a particular treatment and their bacterial numbers (in millions per ml) were measured at a particular age (in days).

Find the rank correlation (Spearman) coefficient.

Age	1	1	2	2	2	2	3
Bacterial No.	336	242	1058	1014	648	1048	1348

(cont.)

Age	7	7	7	14	14	14	16
Bacterial No.	2072	2925	2240	2825	2560	4900	3550

Question 4 (attempt either 4.1=25 marks or 4.2 = 25 marks)

4.1 (25 marks)

A group of 10 strawberry plants is grown in ground treated with a chemical soil conditioner, and the mean yield per plant is 114 g. Experience has shown that when the same variety of strawberry is grown under similar conditions, but with no soil conditioner, the mean has been 110 g and the variance 84.

Test whether it can reasonably be claimed that the soil conditioner had a beneficial effect on yield.

U5

4.2 (25 marks)

Seven plants of wheat grown in pots and given a standard fertilizer treatment yield respectively 8.4, 4.5, 3.8, 6.1, 4.7, 11.2 and 9.6 g dry weight of seed. A further eight plants from the same source are grown in similar conditions but with a different fertiliser and yield respectively 11.6, 7.5, 10.4, 8.4, 13.0, 7.0, 9.6, 13.2 g.

Test whether the two fertilizer treatments have different effects on seed production.

STATISTICAL FORMULAE FOR AEM 201 EXAM PAPER ,
Lecturer Dr Gabriel TATI

MID - POINT

DEFINITION

a) The mid-point of a class is defined as that point lying mid-way between the two class boundaries. It is calculated as

$$\frac{l.c.b + u.c.b}{2}$$

a) A frequency polygon for a frequency distribution having equal class intervals is formed by plotting (as points) class frequencies above the mid-points of the classes to which they relate and joining these points using straight lines.

MEAN

DEFINITION 1

The arithmetic mean (or just mean) of a set $\{x_1, x_2, x_3, \ldots, x_n\}$ is denoted by $\bar{x}$

(and read as 'x bar') and defined as :

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \ldots + x_n) = \frac{1}{n}\sum_{i}^{n} x_i$$

(i.e. $\bar{x}$ is the sum of the items divided by the number of items)

DEFINITION 2

For a discrete frequency distribution taking values $(x_1, x_2, \ldots, x_n)$ with corresponding frequencies $(f_1, f_2, \ldots, f_n)$, the mean $\bar{x}$ is given by

$$\bar{x} = \sum_{i=1}^{n} f_i x_i / \sum_{i=1}^{n} f_i$$

MEDIAN

DEFINITION 1

For a discrete frequency distribution taking the values $(x_1, x_2, \ldots, x_n)$ with corresponding frequencies $(f_1, f_2, \ldots f_n)$, the median is the $\frac{N+1}{2}$-th value when the values are ranked, where $N = \sum f$.

DEFINITION 2

Given a continuous (or grouped discrete) frequency distribution, having determined the median class, and estimate of the median is given by

$$m = L + \left[\frac{\left(\frac{N}{2} - f_L\right)}{f} \right] * C$$

where
L = l.c.b of median class
N= total number of items $\left(=\sum_{i=1}^{n} f_i\right)$
f_L = cumulative frequency up to point L
f=median class frequency
c= median class length

MODE

DEFINITION 1

The Mode of a set of values is defined as that one which occurs with the greatest frequency

Note that for a set that has no repeated values a mode will not exist.

48

DEFINITION 2

For a continuous (or grouped discrete) frequency distribution, given the modal class, an estimate of the mode is given by:

$$L + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \right]$$

where: L= l.c.b. of modal class

$\Delta 1$ = difference in frequencies between modal class and previous class
$\Delta 2$ = difference in frequencies between modal class and following class

c = width of modal class

Note that the quantity $\frac{\Delta_1}{\Delta_1 + \Delta_2}$ is always strictly between 0 and 1 ensuring that the mode must lie in the pre-defined modal class.

STANDARD DEVIATION AND VARIANCE

DEFINITION

The standard deviation of a set of numbers ($x_1, x_2, \ldots, x_n$) with mean $\bar{x}$ is denoted by s and defined:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

In words, s is the square root of the mean of the squares of deviation from the eman (and, hence, is sometimes called the root mean square deviation)

Ignoring the square root sign, we have

The standard deviation of a set of numbers ($x_1, x_2, x_3, \ldots, x_n$) can be expressed using the computational formula

$$s = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2}$$

FOR A FREQUENCY DISTRIBUTION

DEFINITION

For a discrete frequency distribution, the standard deviation is defined:

$$s = \sqrt{\frac{\Sigma f_i \, (x_i - \bar{x})^2}{\Sigma f_i}}$$

and written (for computational purposes) as:

$$s = \sqrt{\frac{\Sigma f_i x_i^2}{\Sigma f_i} - \left[\frac{\Sigma f_i x_i}{\Sigma f_i} \right]^2}$$

Note that $\dfrac{\Sigma f_i \, x_i}{\Sigma f_i}$

is the mean $\bar{x}$ of a frequency distribution

VARIANCE

DEFINITION

The variance of a set, or distribution, of numbers is defined as the square of the standard deviation and is denoted (in an obvious way) by s^2

Precise expressions for the variance in particular situations are given by:

a) For a set

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n} \quad \text{(definition)}$$

$$= \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 \quad \text{(computational formula)}$$

b) For a frequency distribution

$$s^2 = \frac{\Sigma f (x - \bar{x})^2}{\Sigma f} \quad \text{(definition)}$$

50

$$s^2 = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{f}\right)^2$$

(computational formula)

(c) Coding, using $X = \dfrac{x - a}{b}$

$s^2 = b^2 S^2$ (for all data)

SKEWNESS AND KURTOSIS

Skewness is a measure of non-symmetry. For distributions that are skewed to the left or right, the approximate relation between the mode, the median and the mean

Mode =Mean – 3 (Mean – Median))

For a non- skewed distribution (symmetric), it can be empirically shown that the three main measures coincide.

PROBABILITY

DEFINITION

Let an experiment have an outcome set S with E as any event. We define the probability of E occurring, written as Pr (E) or P(E) or P{E}, as a number satisfying the following conditions:

(a) $0 \leq \Pr(E) \leq 1$.

(b) $\Pr(S) = 1$

(c) If E_1 and E_2 are two mutually exclusive events of S, then:

$\Pr(E_1 \text{ or } E_2) = \Pr(E_1) + \Pr(E_2)$ $\Rightarrow \Pr(E_1 \text{ and } E_2) = 0$

(d) If $E_1, E_2, E_3, \ldots, E_n$ are n mutually exclusive events of S, then :

$\Pr(E_1 \text{ or } E_2 \text{ or } E_3 \text{ or } \ldots \text{ or } E_n) = \Pr(E_1) + \Pr(E_2) + \Pr(E_3) + \ldots + \Pr(E_n)$

The above sections are relevant to all types of outcome set, finite or infinite.

(e) If the outcome set S is finite with exactly n outcomes $s_1, s_2, \ldots, s_3$ say, then

$\Pr(s_1) + \Pr(s_2) + \ldots + \Pr(s_n) = 1$

5

i.e $$\sum_{i=1}^{n} Pr(s_i) = 1$$

If the outcome set S is finite with eually likely outcomes (that is, the probability of occurrence of each of them is the same), then the probability of event E occurring is given by:

$$Pr(E) = \frac{n(E)}{n(S)}$$

STATEMENT

If E1 and E2 are any two events of the same experiment, then

$$Pr(E_1 \text{ or } E_2) = Pr(E_1) + Pr(E_2) - Pr(E_1 \text{ and } E_2)$$

Or $$Pr(E_1 \cup E_2) = Pr(E_1) + Pr(E_2) - Pr(E_1 \cap E_2)$$

CONDITIONAL PROBABILITY AND INDEPENDENCE

DEFINITION

Let E_1 and E_2 be any two events, not necessarily from the same experiment. The conditional probability of E1, given E2 has occurred, is written as $Pr(E_1/E_2)$ and defined:

$$Pr(E_1/E_2) = \frac{Pr(E_1 \text{ and } E_2)}{Pr(E_2)}$$

This definition is applicable to all types of outcome sets both finite and infinite.

If S is a finite, equally likely outcome set of some experiment and E_1, E_2 are any two events of S, then:

(a) $$Pr(E_1/E_2) = \frac{n(E_1 \text{ and } E_2)}{n(E_2)}$$

(b) $Pr(\bar{E}_1 / E_2) = 1 - Pr(E_1/E_2)$

(a) If E_1 and E_2 are mutually exclusive, $Pr(E_1/E_2) = 0$

5

DEFINITION

(a) Two experiments are independent if the result of one can in no way effect the possible result of the other

(b) Two events (E1 and E2, say) are independent if the probability that one of them occurs is in no way influenced by whether or not the other has occurred.

This enables us to write $Pr(E1) = Pr(E1/E2) = Pr(E1/\overline{E2})$, and similarly $Pr(E2)=Pr(E2 / E1)$

STATEMENT

If E1 and E2 are any two events::

(a) $Pr(E1 \text{ and } E2) = Pr(E1)*Pr(E1/E2)$
$= Pr(E1)*Pr(E2/E1)$

(b) $Pr(E1 \text{ and } E2) = Pr(E1)*Pr(E2)$, if and only if E1, E2 are independent.

REGRESSION AND CORRELATION

THE LEAST SQUARES REGRESSION

The least square regression of Y on X is that line $Y = b + aX$

A computational expression of ***a*** is given by :

$$a = \frac{n\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

and ***b*** is given by:

$b = \overline{Y} - a*\overline{X}$

CORRELATION

The correlation coefficient, r , is defined in terms of the formula

$$r = \frac{\Sigma\,(X - \bar{X})\,(Y - \bar{Y})}{\sqrt{[\Sigma\,(X - \bar{X})^2][\,\Sigma(Y - \bar{Y})^2]}}$$

We can introduce a computing formula for r that involves the five sums previously obtained in connection with the computations of a and b. The formula is

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

The correlation coefficient is also named Product moment correlation coefficient.

(Product-moment) Correlation coefficient interpretation, r

Rules of thumb for different ranges of r

- 0.75 – 0.99 high degree of relatedness
- 0.50 - 0.74 moderate degree of correlation
- 0.25 - 0.49 low or weak degree of correlation
- below 0.25 very weak degree of association

TEST OF DIFFERENCE BETWEEN MEANS

Construction of the Test

Testing hypotheses about two population means for small samples $N1 \angle 30$, $N2 \angle 30$

Ho: $\mu_1 = \mu_2$

H1: $\mu 1$ # $\mu 2$ (two-tailed test)

Or in situation of

directional alternative hypothesis (one-tailed test)

$H_1: \ \mu_1 < \mu_2$ or $H_1: \ \mu_1 > \mu_2$

If X_1 and X_2 are independent, normally distributed random variables with unknown variances that are assumed to be equal, the appropriate statistic to test that $\mu 1 = \mu 2$ (Ho) is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

With $$S^2 = \frac{[\Sigma(X - \overline{X}_1)^2]_{(1)} + [\Sigma(X - \overline{X}_2)^2]_{(2)}}{(N_1 + N_2 - 2)} = \frac{[\Sigma X^2 - N_1\overline{X}^2]_{(1)} + [\Sigma X^2 - N_2\overline{X}^2]_{(2)}}{(N_1 + N_2 - 2)}$$

(1) refers to the first sample of size N_1, and (2) refers to the second sample of size N_2

The test compares the value of t above to that of t_α (read from the table) at the given level of significance α and degree of freedom $df = N_1 + N_2 - 2$

QUARTILES

For a grouped frequency distribution, given the 1st and 3rd quartile classes, an estimate of Q_1 and Q_3 using the method of interpolation is given by :

$$Q_1 = L_1 + \left[\frac{\frac{n}{4} - f_{L1}}{f_1} \right] * c_1$$

and

$$Q_3 = L_3 + \left[\frac{\frac{3n}{4} - f_{L3}}{f_3} \right] * c_3$$

Where: L_1 and L_3 are the lower class boundaries of the 1st and 3rd quartile classes respectively;

n is the total number of items in the distribution;

f_{L1} and f_{L3} are the cumulative frequencies up to the respective lower bounds L_1 and L_3;

f_1 and f_3 are the frequencies of the 1st and 3rd quartile classes respectively; and

c_1 and c_3 are the widths of the 1st and 3rd quartile classes respectively

Spearman's rank correlation

Given a set of bivariate data (x_1, y_1) (x_2, y_2) ..... (x_n, y_n)

The spearman's rank correlation coefficient is calculated by:

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

Where d is the difference of corresponding ranked pairs of x and y values.

APPENDIX F Distribution of *t*

	Level of significance for one-tailed test					
df	*.10*	*.05*	*.025*	*.01*	*.005*	*.0005*
	Level of significance for two-tailed test					
	.20	*.10*	*.05*	*.02*	*.01*	*.001*
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291