UNIVERSITY OF SWAZILAND

MAIN EXAMINATION PAPER 2016

TITLE OF PAPER	:	INFERENTIAL STATISTICS
COURSE CODE	:	ST 220
TIME ALLOWED	:	THREE (3) HOURS
REQUIREMENTS	:	CALCULATOR AND STATISTICAL TABLES
INSTRUCTIONS	:	THIS PAPER HAS SIX (6) QUESTIONS AND TWO SECTIONS. ANSWER ALL QUESTIONS IN SECTION ONE, ANY THREE (3) QUESTIONS IN SECTION TWO

SECTION ONE

(ANSWER ALL QUESTIONS)

Question 1 [10 marks, 1 mark each]

Choose the correct answer from the alternatives provided.

1. Which of the following is not a property of the normal distribution?

   (a) It is symmetric about its mean

   (b) It is bell-shaped

   (c) It is common

   (d) It is unimodal

2. The government claims that students earn an average of SZL4500 during their summer break from studies. A random sample of students gave a sample average of SZL3975 and a 95% confidence interval was found to be (SZL3525 $< \mu <$ SZL4425). This interval is interpreted to mean that:

   (a) if the study were to be repeated many times, there is a 95% probability that the true average summer earnings is not SZL4500 as the government claims.

   (b) because our specific confidence interval does not contain the value SZL4500 there is a 95% probability that the true average summer earnings is not SZL4500.

   (c) if we were to repeat our survey many times, then about 95% of all the confidence intervals will contain the value SZL4500.

   (d) if we repeat our survey many times, then about 95% of our confidence intervals will contain the true value of the average earnings of students.

   (e) there is a 95% probability that the true average earnings are between SZL3525 and SZL4425 for all students.

3. Which of the following statements about confidence intervals is *incorrect*?

   (a) If we keep the sample size fixed, the confidence interval gets wider as we increase the confidence coefficient.

   (b) A confidence interval for a mean always contains the sample mean.

   (c) If we keep the confidence coefficient fixed, the confidence interval gets narrower as we increase the sample size.

   (d) If the population standard deviation increases, the confidence interval decreases in width.

   (e) If the confidence intervals for two means do not overlap very much, there is evidence that the two population means are different.

4. A 95 percent confidence interval for the mean time taken to process new insurance policies is (11, 12) days. This interval can be interpreted to mean that:

   (a) only 5 percent of all policies take less than 11 or more than 12 days to process

   (b) only 5 percent of all policies take between 11 and 12 days to process

(c) about 95 out of every 100 such intervals constructed from random samples of the same size will contain the population mean processing time

(d) the probability is .95 that all policies take between 11 and 12 days to process

(e) none of the above

5. A turkey producer knows from previous experience that profits are maximized by selling turkeys when their average weight is 12 kilograms. Before determining whether to put all their full grown turkeys on the market this month, the producer wishes to estimate their mean weight. Prior knowledge indicates that turkey weights have a standard deviation of around 1.5 kilograms. The number of turkeys that must be sampled in order to estimate their true mean weight to within 0.5 kilograms with 95% confidence is:

   (a) 35
   (b) 5
   (c) 65
   (d) 10
   (e) 150

6. A confidence statement includes what two things?

   (a) margin of error and bias
   (b) bias and variability
   (c) bias and confidence level
   (d) confidence level and margin of error

7. I read an advertisement recently in which a credit card company promised that I could reduce my debt by 150 percent. Which of the following statements is (are) true?

   (a) This is possible if my debt is more than 150 dollars.
   (b) This is possible if my debt has recently increased by at least 150 percent.
   (c) The company's claim makes no sense.
   (d) Both (a) and (b).

8. The alternative hypothesis for the Chi-square test of independence is that the variables are

   (a) dependent
   (b) related
   (c) independent
   (d) always zero

9. The diameter of ball bearings are known to be normally distributed with unknown mean and variance. A random sample of size 25 gave a mean 2.5 cm. The 95% confidence interval had length 4 cm. Then

   (a) The sample variance is 4.86.
   (b) The sample variance is 26.03.

(c) The population variance is 4.84.

(d) The population variance is 23.47.

(e) The sample variance is 23.47.

10. The 0.01 level of significance is used in an experiment and a two-tailed hypothesis test applied. Computed z is found to be -2.0. This indicates:

    (a) H_0 should be accepted

    (b) We should reject H_0 and accept H_1

    (c) We should have used the 0.05 level of significance.

    (d) None of these is correct.

Question 2 [15 marks, 1 mark each]

State whether each of these statements is true or false, giving brief reasons why this is so (*Note that no marks will be awarded for a simple true/false reply*)

1. When using a large random sample, we cannot assume that its mean forms part of a normal distribution.
2. The least squares regression line minimizes the sum of absolute deviations.
3. The power of a test is the probability of a type 2 error.
4. If two variables are correlated then they must have a linear relationship.
5. The sampling distribution of the mean is distributed the same way as the original observations.
6. A chi-squared value can be positive.
7. The significance level of a test is greater than the probability of a Type 1 error.
8. The mean of a dataset is always smaller than the mode.
9. Two 95% confidence intervals for the same parameter should have the same length.
10. If $\mathfrak{P}(A|B) > \mathfrak{P}(B|A)$ then $\mathfrak{P}(A) > \mathfrak{P}(B)$.
11. If H_0 is false, a high level of power increases the probability we will reject it.
12. Other things being equal, larger confidence levels provide a smaller margin of error.
13. The critical region for rejection of H_0 is the area under the curve which contains all the values of the statistic which fail to allow rejection of H_0.
14. Confidence statements are statements applicable only to the sample of individuals measured.
15. Nonsampling errors can not occur in a census.

SECTION TWO

(ANSWER ANY THREE QUESTIONS)

Question 3 [25 marks, 8+6+8+3]

It is assumed that there is a linear relationship between yield of apple trees and the amount of fertiliser supplied to them. In order to test this assumption, nine apple trees of the same type were randomly selected and supplied weekly with a fixed quantity (x grams) of fertiliser. The yield of each apple tree (y kilograms) was recorded.

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y	3.9	4.3	5.5	6.4	6.9	7.1	7.3	7.7	8.0

(a) Calculate the least squares line of y on x ($\sum x^2 = 96$, $\sum xy = 186.75$ and $\sum y^2 = 379.51$).

(b) Compute and interpret the coefficient of determination.

(c) Is the *relationship* between the *amount of fertiliser* supplied and *yield* meaningful (or significant)? Use $\alpha = 0.05$.

(d) What prediction will you give for a tree that is treated weekly with 0.0032 kg of fertiliser?

Question 4 [20 marks, 10+15]

(a) It is suspected that the number of faulty products produced by a manufacturing industry varies depending on which day of the week they were made. In order to test this a random sample of 200 products is taken from the warehouse, and after inspection yields the following results.

	Day of manufacture		
Product	**Monday**	**Tuesday-Thursday**	**Friday**
Perfect	32	94	34
Faulty	3	21	16

Use the χ^2 test to decide if these results indicate that there is a connection between faulty products and when they were made.

(b) The production manager of Raylite batteries, a car battery manufacturer, wants to know whether the three machines used for this process (labelled A, B and C) produce equal amount of rejects. A random sample of shifts for each machine was selected and the number of rejects produced per shift was recorded. The number of shifts selected were 6, 4 and 5 for machines A, B and C respectively. The average number of rejects for machines A and C were 10.5 and 14.0 respectively. The following information is provided as well.

	Degrees of freedom	Sum of squares	Mean sum of squares	F value
Battery				4.165
Residuals		76.5		
Total				

By completing the table using the information provided, can the production manager of Raylite batteries concluded that the three machines used to manufacture car batteries produce rejects at the *same average rate* per shift. Use $\alpha = 0.05$ and show the ANOVA table. Also state the necessary assumptions.

Question 5 **[25 marks, 10+10+5]**

(a) A consumer report examined potential differences between two brands of tyres. The mean life of the tyres is of primary concern. The available data, measured in thousands of miles, are provided below:

	Sample size n	**Sample mean m**	**Sample standard deviation s**
Brand A	34	21.4	1.5
Brand B	38	22.3	1.8

Assuming that the population variances for the two *brands* are equal, use an appropriate hypothesis test to determine whether the mean lives of the two brands are different.

(b) The student union of a large university gathered a random sample of 525 students to determine whether they are in favour of a new grading system. The results are summarized in the table below.

	Sample size	**Number in favour or new grading system**
Humanities	325	221
Science	200	120

(i) Do the results indicate a difference between humanities and science in the population proportions in favour of the new grading system? Conduct an appropriate test and comment of your results.

(ii) Give a 97% confidence interval for the difference between the two proportions in the population.

Question 6 **[25 marks, 10+4+6+5]**

(a) Ten randomly selected oil wells in a large field of oil wells produced 21, 19, 20, 22, 24, 21, 19, 22, 22, and 20 barrels of crude oil per day. Is this evidence at the 0.01 level of significance that the oil wells are not producing an average of 22.5 barrels of crude oil per day?

(b) Suppose the probability is 0.30 that any given student in a large class can provide the answer to an assigned problem. What is the probability that the fourth student randomly selected by the instructor will be the first one who can provide the answer to the problem?

(c) A test is taken by some students, their marks are recorded and we are interested in the properties of the sample mean. Under the assumption that the marks follow a Normal distribution with exact mean 60 and variance 81, calculate the probability that the mark of a randomly selected student?

(i) is greater than 59.5 exactly; and

(ii) lies between 59 and 60.5 exactly.

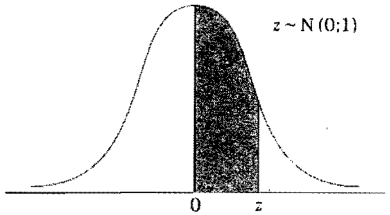
(d) A blended wine is intended to comprise two parts of Sauvignon to one part of Merlot. The amounts dispensed to make up a nominal 75cl bottle of this wine are X cl of Sauvignon and Y cl of Merlot, where X and Y are assumed to be independent Normally distributed random variables with respective means 52 and 26 cl and respective variances 1 and 0.5625. Find the probability that the actual volume of wine dispensed into a bottle is less than the nominal volume.

APPENDIX 1: LIST OF STATISTICAL TABLES

TABLE 1

The standard normal distribution (*z*)

This table gives the area under the standard normal curve between 0 and *z* i.e. $P[0 < Z < z]$

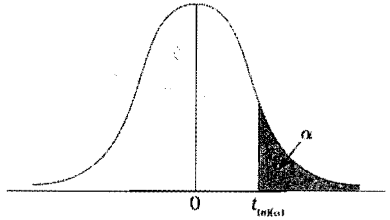


Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

TABLE 2

The *t* distribution

This table gives the value of $t_{(n)(\alpha)}$ where *n* is the degrees of freedom i.e. ▇ = $P[t \geq t_{(n)(\alpha)}]$

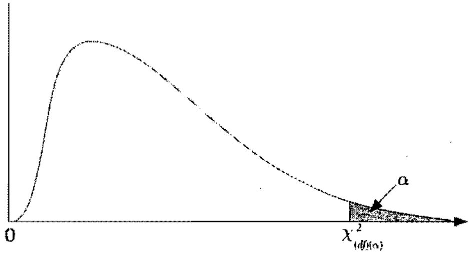


n	0.100	0.050	0.025	0.010	0.005	0.0025
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

TABLE 3

The Chi-Squared distribution (χ^2)

This table gives the value of $\chi^2_{(df)(\alpha)}$ where df is the degrees of freedom i.e. $\alpha = P[\chi^2 > \chi^2_{(df)(\alpha)}]$

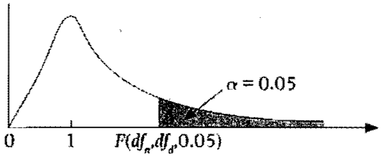


α	0.100	0.050	0.025	0.01	0.005	0.0025
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

TABLE 4 (a)

F distribution ($\alpha = 0.05$)

The entries in this table are critical values of F for which the area under the curve to the right is equal to 0.05.



Degrees of freedom for denominator	Degrees of freedom for numerator
[illegible]	[illegible]

TABLE 4 (a) continued

F distribution ($\alpha = 0.05$)

Degrees of freedom for denominator	Degrees of freedom for numerator
[illegible]	[illegible]

APPENDIX 2: LIST OF KEY FORMULAE

MEASURES OF CENTRAL LOCATION

Arithmetic mean

Ungrouped data

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

Grouped data

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

Mode

Grouped data

$$M_o = O_{mo} + \frac{c(f_m - f_{m-1})}{2f_m - f_{m-1} - f_{m+1}}$$

Median

Grouped data

$$M_e = O_{me} + \frac{c\left[\frac{n}{2} - f(<)\right]}{f_{me}}$$

Lower quartile

Grouped data

$$Q_1 = O_{q1} + \frac{c\left(\frac{n}{4} - f(<)\right)}{f_{q1}}$$

Upper quartile

Grouped data

$$Q_3 = O_{q3} + \frac{c\left(\frac{3n}{4} - f(<)\right)}{f_{q3}}$$

Geometric mean

Ungrouped data

$$GM = \sqrt[n]{x_1 \times x_2 \times x_3 \times \ldots \times x_n}$$

Weighted arithmetic mean

Grouped data

$$\text{weighted } \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

MEASURES OF DISPERSION AND SKEWNESS

Range Range = Maximum value – Minimum value + 1 $= x_{max} - x_{min} + 1$

Variance *Mathematical – ungrouped data*

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{(n-1)}$$

Computational – ungrouped data

$$s^2 = \frac{\Sigma x_i^2 - n\bar{x}^2}{(n-1)}$$

Standard deviation $s = \sqrt{s^2}$

Coefficient of variation $CV = \frac{s}{\bar{x}} \times 100\%$

Pearson's coefficient of skewness

$$sk_p = \frac{n\Sigma(x_i - \bar{x})^3}{(n-1)(n-2)s^3}$$

$$sk_p = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$ (approximation)

PROBABILITY CONCEPTS

Conditional probability $P(A/B) = \frac{P(A \cap B)}{P(B)}$

Addition rule *Non-mutually exclusive events*
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Mutually exclusive events
$P(A \cup B) = P(A) + P(B)$

Multiplication rule *Statistically dependent events*
$P(A \cap B) = P(A/B) \times P(B)$

Statistically independent events
$P(A \cap B) = P(A) \times P(B)$

***n*! = *n* factorial** $n \times (n-1) \times (n-2) \times (n-3) \times \ldots \times 3 \times 2 \times 1$

Permutations ${}_nP_r = \frac{n!}{(n-r)!}$

Combinations ${}_nC_r = \frac{n!}{r!\,(n-r)!}$

PROBABILITY DISTRIBUTIONS

Binomial distribution $P(x) = {}_nC_x\, p^x (1-p)^{(n-x)}$ for $x = 0, 1, 2, 3, \ldots, n$

$P(x \text{ successes}) = \frac{n!}{x!\,(n-x)!}\, p^x (1-p)^{(n-x)}$ for $x = 0, 1, 2, 3, \ldots, n$

Binomial descriptive measures
Mean $\mu = np$
Standard deviation $\sigma = \sqrt{np(1-p)}$

Poisson distribution $P(x) = \frac{e^{-a} a^x}{x!}$ for $x = 0, 1, 2, 3 \ldots$

Poisson descriptive measures
Mean $\mu = a$
Standard deviation $\sigma = \sqrt{a}$

Standard normal probability $z = \frac{x - \mu}{\sigma}$

12

CONFIDENCE INTERVALS

Single mean

n large; variance known

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z\frac{\sigma}{\sqrt{n}}$$

(lower limit) (upper limit)

n small; variance unknown

$$\bar{x} - t_{(n-1)}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{(n-1)}\frac{s}{\sqrt{n}}$$

(lower limit) (upper limit)

Single proportion

$$p - z\sqrt{\frac{p(1-p)}{n}} \le \pi \le p + z\sqrt{\frac{p(1-p)}{n}}$$

(lower limit) (upper limit)

HYPOTHESES TESTS

Single mean

Variance known

$$z\text{-}stat = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Variance unknown; n small

$$t\text{-}stat = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$ 8.2

Single proportion

$$t\text{-}stat = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Difference between two means

Variances known

$$z\text{-}stat = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ 9.1

Variances unknown; n_1 and n_2 small

$$t\text{-}stat = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$ 2

Paired *t*-test

$$t\text{-}stat = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

where $\mu_d = (\mu_1 - \mu_2)$

and $s_d = \sqrt{\frac{\Sigma(x_d - \bar{x}_d)^2}{n-1}}$

Differences between two proportions

$$z\text{-}stat = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } \hat{\pi} = \frac{x_1 + x_2}{n_1 + n_2};\ p_1 = \frac{x_1}{n_1};\ p_2 = \frac{x_2}{n_2}$$ 9.8

Chi-Squared

$$\chi^2\text{-}stat = \Sigma\frac{(f_o - f_e)^2}{f_e}$$

Overall mean

$$\bar{\bar{x}} = \frac{\Sigma\Sigma x_{ij}}{N}$$ 11

Total sum of squares (SSTotal) $= \sum_i\sum_j (x_{ij} - \bar{\bar{x}})^2$

SST $= \sum_j^k n_j(\bar{x}_j - \bar{\bar{x}})^2$

SSE $= \sum_j\sum_i (x_{ij} - \bar{x}_j)^2$

SSTotal $=$ SST + SSE

MSTotal $= \frac{\text{SSTotal}}{N-1}$

MST $= \frac{\text{SST}}{k-1}$

MSE $= \frac{\text{SSE}}{N-k}$

F-stat $= \frac{\text{MST}}{\text{MSE}}$