THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF APPLIED MATHEMATICS

Subject Code: AMA1501/ AMA1602

Subject Title: Introduction to Statistics for Business/ Introduction to Statistics

Session: Semester 2, 2022/2023

Date: 20 Apr 2023

Time: 15:15 – 18:15

Time Allowed: THREE Hours

This question paper has 14 pages (attachments included).

Instructions to Candidates: This question paper has 6 questions.

Attempt any **FIVE** questions.

Each question carries equal marks.

Attachments: Formula Sheets, Standard Normal Distribution Table, Student's t-distribution Table, χ^2 Distribution Table and F-distribution Table

DO NOT TURN OVER THE PAGE UNTIL YOU ARE TOLD TO DO SO

Attempt any **FIVE** questions.

1. In order to study the age distribution of certified accountants in Hong Kong in 2023, 100 accountants are randomly selected and interviewed. The results are given in the following table.

Age	Number of accountants
20 to 29	7
30 to 34	15
35 to 39	18
40 to 44	12
45 to 49	12
50 to 54	13
55 to 59	11
60 to 79	12

Source: The Hong Kong Institute of CPAs

(a) Calculate the mean, mode, standard deviation and semi-interquartile range of the age distribution. [10 marks]

(b) Calculate the coefficient of skewness using the results in (a). Comment the skewness of the age distribution. [2 marks]

(c) Estimate, from the frequency distribution table, the proportion of accountants aged between 38 to 48 years. [4 marks]

(d) Construct a 99% confidence interval for the mean age of all accountants. [4 marks]

2. (a) Suppose you are organizing a charity event and you have 20 volunteers who have signed up to help. You need to divide them into 4 teams, each with 5 volunteers. How many different ways can this be done? [4 marks]

   (b) If Emily flips an unfair coin twice and the probability of getting two heads is 0.16. If the same coin is flipped again twice, what is the probability of getting two tails? [4 marks]

   (c) A student is taking a test which consists of 3 questions. The probabilities that he answers each of the questions correctly are 0.3, 0.4 and 0.5 respectively. Find the probability that he answers exactly two questions correctly in the test. [4 marks]

   (d) Suppose there are four boxes on a game show that a participant can choose from, namely the red box with one \$100 bill and nine \$1 bills, the green box with two \$100 bills and eight \$1 bills, the blue box with three \$100 bills and seven \$1 bills, and the yellow box with five \$100 bills and five \$1 bills. If a participant randomly selects a box and then randomly picks a bill from that box, what is the probability that the \$100 bill was taken from the yellow box, provided that a \$100 bill was selected? [8 marks]

3. (a) A statistics test was taken by students in classes A and B. The number of students in each class is equal and their test scores are assumed to follow a normal distribution. The passing mark for the test is 45. In class A, 99.7% of students scored higher than 29 and 75.8% of students scored lower than 42.8. In class B, the average score is 42 and the standard deviation is 5.

      i. Find the proportions of passing students in each class. Which class has less students passing the test? [8 marks]

      ii. In class B, what is the probability that the 5th randomly chosen student is the 3rd one who passes the test? [4 marks]

      iii. Students in both classes who have scored k marks or less should attend a remedial program. Suppose there are 500 students in each class. What should the greatest value of k be so that we can expect at most 51 students in class B will join the remedial program? Give your answer as an integer. [4 marks]

(b) Given that a laboratory uses a piece of equipment on average 1.3 times per day and assuming that the usage of the equipment per day follows a Poisson distribution, what are the probabilities of the following events occurring?

i. The equipment is used less than 2 times in a day. [2 marks]

ii. The equipment is used less than 2 times in each of two successive days. [2 marks]

4. (a) The standard deviation of the travelling time taken for a shuttle bus to finish a trip is 2.3 minutes. Based on a random sample of 36 trips, the sample mean of the travelling time taken to finish a trip is 15 minutes. Construct a 90% confidence interval for the population mean of the travelling time taken for the shuttle bus to finish a trip. [4 marks]

(b) Assume that the amounts spent by customers in a restaurant follow a normal distribution with a standard deviation of \$15. The following are the amounts spent (in \$) by 10 customers selected randomly and independently from the restaurant.

84 88 90 92 92 95 100 107 110 120

Construct a 99% confidence interval for the population mean of the amounts spent by customers. [6 marks]

(c) In a factory, jam is packed into tins labelled with weights of 800 g each. The manager of the factory would like to estimate the population proportion p of tins of jam with weights over 800 g. Thus, he selected 150 tins of jam randomly and independently, and found that only 96 of them are actually over 800 g. Construct a 95% confidence interval for p. [4 marks]

(d) All applicants for a specific position are required to take an aptitude test as part of the screening process. Suppose that the test has a mean score of 75 and a standard deviation of 20. Assume the distribution of scores is approximately normal.

i. In a group of 100 applicants, how many would you expect to score below 60? [3 marks]

ii. What is the probability that the mean of a group of 100 applicants will score below 70? [3 marks]

5. (a) A study is being conducted to compare the effectiveness of two types of medication for a certain disease in adult patients. The goal is to determine if there is a significant difference in the percentage of patients experiencing adverse reactions. In a random sample of 200 adults given medication A, 20 patients still had the disease 30 minutes after taking the medication. In another random sample of 200 adults given medication B, 12 patients still had the disease 30 minutes after taking the medication. The significance level for the test is set at the 1%. [6 marks]

(b) The table below presents the non-institutional population residing on land across two District Council districts and age groups, as of the year 2022.

	Age group (in thousands)			
	0 – 14	15 – 24	25 – 64	65 and over
Central and Western	22.6	15.7	138.3	44.2
Wan Chai	15.6	9.4	96.6	34.9

Source: Census and Statistics Department, HKSAR

Test whether the two attributes are independent at the 10% level of significance. [8 marks]

(c) As part of a food research conducted to investigate the impact of a recently developed type of baby food, we documented the weight (measured in pounds) of 8 babies before and after they consumed the food for a week, as presented below.

Baby	1	2	3	4	5	6	7	8
Before consumption	49	53	51	52	47	50	52	53
After consumption	52	55	52	53	50	54	54	53

Determine whether the weight gain observed in the babies is significant or not at the 5% level of significance. State any assumption(s)/approximation(s) used. [6 marks]

6. The following table presents data on the number of supermarket establishments and their respective employee counts in Hong Kong, spanning from 2015 to 2021.

Year	Number of establishments	Number of employees (in thousands)
2015	75	29.2
2016	70	28.6
2017	77	28.5
2018	83	27.9
2019	79	27.4
2020	94	27.2
2021	115	27.3

Source: Census and Statistics Department, HKSAR

A regression model is developed based on the above data.

(a) Calculate the rank correlation coefficient, r_s, for the data. What does the result imply about the relationship between the two variables? [4 marks]

(b) Find the least squares equation for predicting the number of employees based on the number of establishments. [6 marks]

(c) Estimate the number of employees that can be expected if the number of establishments reaches 80 in a given year. [2 marks]

(d) With the coefficient of determination $r^2 = 48.7365\%$, give the missing values denoted by variables **a** to **f** in the ANOVA table provided below. [4 marks]

ANOVA Table

Source	*df*	*SS*	*MS*	*F*
Regression	**a**	**b**	**d**	**f**
Residual	5	**c**	**e**	
Total	6	3.5486		

(e) Test the significance of the regression model at the 5% level of significance. [4 marks]

End

Formula sheet

1. Sample Statistics:

	Ungrouped data	Grouped data
Arithmetic Mean	$\dfrac{\Sigma x}{n}$	$\dfrac{\Sigma fx}{\Sigma f}$
Standard Deviation	$\sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\dfrac{\Sigma x^2 - (\Sigma x)^2/n}{n-1}}$	$\sqrt{\dfrac{\Sigma f(x-\bar{x})^2}{\Sigma f - 1}} = \sqrt{\dfrac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f - 1}}$

2. Probability Distributions:

   (a) Binomial $P(r) = {}_nC_r p^r (1-p)^{n-r}$

   (b) Poisson $P(r) = \dfrac{e^{-\lambda}\lambda^r}{r!}$

3. Standard Errors:

   (a) Mean $\dfrac{\sigma}{\sqrt{n}}$

   (b) Proportion $\sqrt{\dfrac{p(1-p)}{n}}$

   (c) Difference between means $\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

   (d) Difference between proportions $\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$

4. Test Statistics:

   (a) $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ (one sample)

   $$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \quad \text{(two samples)}$$

(b) $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ (one sample)

$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ (two samples) where $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

(c) $\chi^2 = \sum \dfrac{(O - E)^2}{E}$

5. Correlation and Regression:

(a) Product moment correlation coefficient

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

(b) Spearman's rank correlation coefficient

$$R_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

(c) Least squares regression line $y = a + bx$

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \qquad a = \frac{\Sigma y}{n} - \frac{b\Sigma x}{n}$$