**STAT S251F**
**Statistical Data Analysis**
**Autumn Term, 2020**
**Assignment 1**

---

Student Name: ______________________________

Student ID: ______________________________ Mark: ________________________

---

Points to Note:

(1) ***Submission Deadline: ≤ 11:59pm, 20 November 2020 (Please submit through OLE)***

(2) ***Important Note: LATE submission is NOT accepted and no marks will be given.***

(3) ***Students involved in PLAGARISM will receive ZERO marks WITHOUT NOTICE!!!***

(4) ***Use A4 paper to complete your assignment.***

(5) ***Your solutions to the questions MUST be HAND-WRITTEN.***

(6) ***Scan your completed assignment in a SINGLE pdf file.***

(7) ***Submit your pdf file via the OLE. Submission of assignment through email is NOT accepted.***

---

## Section A (10%): Multiple Choices (1 mark @)

***Instruction: Put your answers into the boxes (□) on the right.***

1. OUHK intends to investigate the future study and employment of its students. The researcher randomly selects 1000 students from a total of 12000 students.

   A. 1000 students constitute the population
   B. 12,000 students constitute the sample
   C. 12000 students constitute the population
   D. Population and sample cannot be determined

   □

2. One of the features of random sampling is

   A. The probability of selecting each element must be equal
   B. The sampling results can be valid to the elements in the sample only
   C. Sampling is conducted according to subjective judgment
   D. The sample selected at random is representative of the whole population

   □

3. According to international practice, a population census is conducted every $n$ years. The value of $n$ is

   A. 3
   B. 5
   C. 7
   D. 10

   □

4. The objectives of performing random sampling do not include which of the following?

   A. Use of sample statistics to infer/estimate population parameters
   B. Computation of population mean and population standard deviation, etc.
   C. Collection of respondents' information
   D. The population is inaccessible

   □

5. The mean marks for Statistics of two classes (A and B) are the same, but the standard deviation of marks for Class B is larger. The correct statement of the their academic performance is

   A. Because the mean marks are equal, therefore their academic performance is the same. □

   B. Although the mean marks are equal, however, Class B’s standard deviation is greater, indicating that the potential of learning of the students of Class B is also greater.

   C. The mean marks of the two classes are equal. However, the students in the class with greater standard deviation are unstable.

   D. Since the standard deviations are unequal, therefore the performance is not the same. The students of Class A were not performed stably.

6. Which one of the following is an ordinal variable?

   A. Time □<br>
   B. Distance<br>
   C. Academic qualification<br>
   D. Number of girl friends

7. Examples of ratio data do not include

   A. Temperature □<br>
   B. Intelligent Quotient<br>
   C. The year measured from the birth of Jesus Christ<br>
   D. Time

8. Histograms are only suitable for

   A. Discrete data □<br>
   B. Continuous data<br>
   C. Nominal data<br>
   D. Ordinal data

9. Bar charts are only appropriate for

   A. Quantitative data □<br>
   B. Categorical data<br>
   C. Discrete data<br>
   D. All kinds of data

10. When extreme values exist in a set of data, a suitable tool for measuring central tendency is<br>
   A. Mean<br>
   B. Median<br>
   C. Mode □<br>
   D. Standard deviation

**Section B (10%): True-or-false Questions (1 mark @)**

Put a "✓" if the statement is true; otherwise, put a "×" into the boxes on the right.

---

1. Random sampling methods are commonly adopted by marketing research companies. □

2. A census can be conducted in every situation. □

3. The reliability of survey results has no relationship with the size of the sample used. □

4. Sampling error can be estimated in non-random sampling. □

5. Statistical inference (use sample results to make inference about the population parameters) can be made on the basis of both random and non-random sampling. □

6. Survey sampling is the only way of collecting sample data from a population. □

7. A sample must be representative of a population. □

8. A sampling frame must be comprehensive, exhaustive, and up-to-date. □

9. A sampling unit and an element of a population are identical. □

10. In some survey sampling problems, more than one sampling frame is employed. □

## Section C (80%): Long/Short Questions

*(1) Answer ALL questions in this section.*

*(2) Give your answers in 2 decimal places when the answers cannot be expressed as integers or fractions.*

---

### Question 1 (22 marks)

The Great Ego Group of Companies employed some employees in various departments. Their monthly salaries (measured in '00 HK$) are displayed in the following table:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Administrative | 130 | 130 | 125 | 95 | 83 | 55 | 50 | 50 | 50 | 50 |
| Personnel | 140 | 130 | 90 | 90 | 55 | 55 | 55 | 50 | | |
| Marketing | 78 | 85 | 78 | 50 | 50 | 55 | | | | |

Two officers, Peter and Mary, are asked to compute the mean monthly salary of all the employees in the above three departments. Peter suggests that a one-stage method is appropriate to compute the overall mean monthly salary by using the following formula:

$$\overline{x_{all}} = \frac{1}{24}\sum_{i=1}^{24} x_i$$

where $x_1, x_2, \mathrm{L}\ , x_{24}$ denote the 24 salary data in the above table, ***arranging from left to right and then from top to bottom***.

Alternatively, Mary proposed that a two-stage averaging method is also applicable to calculate the mean. She uses the following mean formulae:

$$\overline{x_{Admin}} = \frac{1}{10}\sum_{i=1}^{10} x_i$$

$$\overline{x_{Per}} = \frac{1}{8}\sum_{i=11}^{18} x_i$$

$$\overline{x_{Mar}} = \frac{1}{6}\sum_{i=19}^{24} x_i$$

*Stage 1*: Compute the means $\overline{x}_{Ad\min}, \overline{x}_{Per}, \overline{x}_{Mar}$ by departments;

*Stage 2*: Compute the overall mean by taking arithmetic mean of $\overline{x}_{Ad\min}, \overline{x}_{Per}, \overline{x}_{Mar}$. That is, the overall mean is computed by the following formula:

$$\overline{\overline{x_{all}}} = \frac{1}{3}(\overline{x_{Admin}} + \overline{x_{Per}} + \overline{x_{Mar}})$$

(a) (i) Compute $\overline{x_{all}}$ using Peter's method. [1]

(ii) Compute $\bar{x}_{Ad\min}, \bar{x}_{Per}, \bar{x}_{Mar}$. Hence, compute $\overline{\overline{x_{all}}}$ by Mary's method. [4]

(iii) Are the results obtained in parts (i) and (ii) consistent? Justify your answer. [9]

(b) Compute the median and mode of the 24 employees. [2]

(c) Among the 3 measures of central tendency: mean, median and mode, you are required to choose a measure from these three listed above to describe the average salary level of the Great Ego Group based on each of the following situation:

(i) If you are the officer-in-charge of the public relations section planning to promote the group's salary level and benefits. [2]

(ii) If you are an employee earning monthly salary of $5,000. [2]

(iii) If you are a tax officer of the Inland Revenue Department (IRD), Hong Kong. [2]

Give reason to your decision in each case.

**Question 2 (17 marks)**

Two sample datasets are collected from a class of students. The first dataset consists of the body weights (kilogram, kg) of the students, whereas the second dataset comprises the times (second, s) the students need to take in running 100 meters. The data are shown below:

Dataset 1 (kg): 45, 48, 57, 46, 59, 61, 44, 42, 56, 60

Dataset 2 (second, s): 9.9, 12.3, 11.5, 13.5, 15.7, 10.8, 11.0, 12.4, 13.6, 10.3

(a) Compute the sample mean and sample standard deviation of each dataset. [4]

(b) Is it appropriate to use sample standard deviation to compare the dispersions between the datasets? Explain your answer briefly. [4]

(c) Your classmate suggests that the formula below is appropriate in comparing their dispersion:

$$\frac{\text{sample standard deviation}}{\text{sample mean}} \times 100\% = \frac{s}{\bar{x}} \times 100\%$$

(i) Use this formula to compute the percentages for the above two samples. [5]

(ii) Justify the appropriateness of his chosen formula for comparison of dispersion between any 2 different datasets. [4]

**Question 3 (18 marks)**

A company consists of 2 departments, A and B only. The mean monthly salary paid to all employees is \$20,000. The respective mean monthly salaries paid to employees of A and B are \$22,000 and \$18,000.

Let $n_A$ and $n_B$ be respectively the number of employees in the departments, A and B.

(a) Determine the ratio of numbers of employees for the 2 departments, $n_A \colon n_B$. [10]

(b) If the total number of employees in the 2 departments is 200, compute the respective numbers of employees in the 2 departments. [2]

(c) The CEO of the company has just established a new department, C. The mean monthly salary and number of employees $(n_C)$ of this third department are \$21,000 and 80. Calculate the <u>overall</u> mean monthly salary for the 3 departments of the company. [6]

**Question 4 (23 marks)**

Suppose that the weights (measured in kg) of pigs of the Kwong Ming Farm are normally distributed with $N(\mu,\ 11.2^2)$. A random sample of size 20 is selected from the whole population and their weights are listed below:

$$158,\ 157,\ 260,\ 162,\ 157,\ 159,\ 160,\ 261,\ 166,\ 167,$$
$$140,\ 141,\ 147,\ 145,\ 259,\ 111,\ 123,\ 234,\ 234,\ 135$$

(a) Estimate the sample mean weight and standard deviation of weights. [4]

(b) Construct a 90% confidence interval for $\mu$. [3]

(c) Under what circumstance can we find the true value of $\mu$? [1]

(d) Among 90% CI, 95% CI, and 99% CI, which one's width is the longest? Explain your answer briefly. [2]

(e) Is it true to say that "The longer the width of a confidence interval, the more reliable the confidence interval is."? Explain your answer briefly. [2]

(f) Could we construct a 100% confidence interval based on a random sample? Explain your answer briefly. [2]

(g) What sample size is needed to have 95% confidence in estimating the population mean height to within $\pm 1kg$? [4]

***The following part (h) is independent of the above parts.***

(h) The following are 2 statements about interval estimation:

(i) $P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$, where only $\sigma$ is known.

(ii) $P(140 < \mu < 160) = 0.95$

Which one of them, (i) or (ii), is a correct statement about confidence interval? Explain your answer briefly. [5]

**[END OF ASSIGNMENT 1]**