# SINGAPORE POLYTECHNIC

## 2020/2021 SEMESTER TWO

## TEST 1

**Specialist Diploma in Data Science (Artificial Intelligence)**
**Specialist Diploma in Data Science (Big Data and Streaming Analytics)**
**Specialist Diploma in Data Science (Data Analytics)**

MS9001 INTRODUCTION TO STATISTICS FOR DATA SCIENCE

Time allowed: 2 hours

Instructions to Candidates:

1. The Singapore Polytechnic examination rules are to be complied with.
2. This paper consists of 4 questions in 10 printed pages.
3. Answer all questions.
4. All answers are to be presented in this paper.
5. Unless otherwise stated, give all non-exact answers to 3 decimal places.
6. You are allowed to make references to materials, as well as use statistical software.
7. Questions 1 and 2 each require a data file which can be downloaded from "Learning Resources" in PolyMall. Please follow your invigilator's instructions.

**Name :** ____________________

**Admin. No. :** ____________________

**Class :** ____________________

| Question No. | Marks |
|---|---|
| 1 (25 marks) | |
| 2 (25 marks) | |
| 3 (25 marks) | |
| 4 (25 marks) | |
| **Total** | |

**Question 1** (25 marks)

In Kentucky, animal bites are often reported to law enforcement (such as animal control). The main concern is the prevention of such occurrences and hence, fines are levied on owners who do not keep a close watch on their pets. A random sample of 280 animal bites in Kentucky was recorded with the following variables: SpeciesIDDesc, GenderIDDesc, Color, WhereBittenIDDesc and Fine (USD). The dataset can be found in '***Animal Bites.xlsx***'. Use it to answer the following questions.

(a) Define the population and sample in this context. (2 mark)

(b) What is the type of data for each of the variables below? (4 marks)

| **Variable** | **Qualitative / Quantitative** | **Nominal / Ordinal / Discrete / Continuous** |
|---|---|---|
| SpeciesIDDesc | | |
| GenderIDDsec | | |
| Color | | |
| WhereBittenIDDesc | | |

(c) Find the 2 measures of centre (i.e. mean and median) of the fines imposed on the owners. (2 marks)

**Mean**: __________

**Median**: __________

(d) Using a histogram to plot out the fines in this dataset, do we have a normal or non-normal distribution? Explain your choice. (2 marks)

(e) By drawing a suitable graph, state which animal species contributed to the bites and write down their percentages. (3 marks)

Animal Species: ____________, ________________, _______________

Percentage of the above animal species:

____________, ________________, _______________respectively.

(f) State which parts of the body were involved in these bites and write down their percentages. (2 marks)

Part of Body: ________________, ____________________

Percentage of the affected body part:

_________________, ________________respectively.

(g) Fill in the blanks below. (Round off the answers to the nearest dollar.) (6 marks)

"The middle 50% of the fines imposed on raccoon owners are between $ _________ and $4,153." IQR is $___________.

" The middle 50% of fines imposed on dog owners are between $1,649 and $___________." IQR is $____________.

"The middle 50% 0f fines imposed on cat owners are between $_____________ and $___________." IQR is $2,287.

(h) Fill in the following descriptive statistics on fines for dog bites. (2 marks)

| Variable | Total Count | Mean | Minimum | Maximum |
|---|---|---|---|---|
| Fine (USD) for dog bites | | | | |

(i) Based on your answers above, which is the most common animal bites in Kentucky? Provide one recommendation on how law enforcement officers can discourage such occurrences. (2 marks)

~~~ Please turn over for next question ~~~
~~~

**Question 2** (25 marks)

The Off-Track Branches of ABC Racing Club must maintain sufficient cash to pay off any claims of winnings by the punters. But if too much cash is unnecessarily kept at the branches, the Club is forgoing the opportunity of investing the money and earning interest, and increasing the risks associated with keeping large amount of cash at the branches.

At a particular branch, the amount of cash being maintained is based on the norm that the average winnings claim over the weekend at $2,000. Perform hypothesis testing to decide if there is evidence to believe that the average winnings claim *has decreased* based on the latest sample obtained given in the dataset (‘***Winnings Claim.xlsx’***), which could mean an opportunity for the branch to reduce the cash to be maintained over the weekend.

(a) Set up the hypothesis by filling in the blanks below: (6 marks)

Let μ be ______________________________________________.

$H_0$: ______________________ $H_1$: ______________________

(b) Provide the following information on the sample data (3 marks)

Sample size, $n$ = ________________________

Sample mean, $\bar{x}$ = $_____________________

Sample SD, $s$ = $_______________________

(c) Use Minitab to find the P-Value and the 95% confidence interval/bound for the mean winnings claim amount. (4 marks)

(d) Interpret the results, using both the P-Value and the 95% confidence interval/bound from part (c) and conclude if there is evidence to believe that the population mean winnings claim is significantly lower than the norm. (8 marks)

(e) What is your answer in (d) if a 90% confidence interval/bound for the mean winnings claim amount is used? Justify your reasoning **without performing another hypothesis test.** (4 marks)

~~~ Please turn over for next question ~~~
~~~

**Question 3** (25 marks)

(a) Decide whether each of the scenarios described below fulfils the conditions of a Binomial distribution by circling the correct answers for all the conditions. (8 marks)

**Scenario 1**

A deck of 52 cards has 4 Kings. Four cards are drawn randomly, one at a time without replacement (i.e. the drawn cards are not put back into the deck).
Let *X* be the number of Kings drawn.

| **Condition** | **Fulfilled?** |
|---|---|
| There is a repeated fixed number of trials or observations, *n*. | Y / N |
| All these trials are independent. | Y / N |
| Each trial should end in one of two outcomes: success or failure. | Y / N |
| The probability of success, *p*, must be the same for all trials. | Y / N |

**Scenario 2**

Mosquito infestation can lead to spread of dengue and other diseases. Inspectors will go around homes to check for mosquito infestation. Proportion of infestation in the general population is 0.03. The inspectors selected 10 homes randomly to check.

Let *X* be the number of homes infested.

| **Condition** | **Fulfilled?** |
|---|---|
| There is a repeated fixed number of trials or observations, *n*. | Y / N |
| All these trials are independent. | Y / N |
| Each trial should end in one of two outcomes: success or failure. | Y / N |
| The probability of success, *p*, must be the same for all trials. | Y / N |

(b) The manufacturer of YOLO buses reported the probability of breakdown on the road of a YOLO bus on any given day is 0.03. LOCO Bus Co Ltd has a fleet of 20 YOLO buses. LOCO will like to ensure that they have enough service staff and spare parts by estimating number of bus breakdowns per day represented by a variable X. Assume X follows a Binomial Distribution.

(i) Define a variable X for the above distribution and write down $X \sim B\ (n, p)$, stating the values of n and p. (2 mark)

(ii) What is the mean number of breakdowns per day? (1 mark)

(iii)What is the probability of getting one breakdown? (2 mark)

(iv)What is the probability of getting at most one breakdown? (2 marks)

(v) What is the probability of getting at least two breakdowns? (2 marks)

(vi)LOCO can hire a permanent staff to service breakdowns or they can outsource. Two breakdowns a day will justify the cost of a breakdown service staff. What should LOCO do?

Justify your answer with reasons from your answers derived above. (3 marks)

(c) The length of time, $Y$ (in hours), of a cell phone that works before it requires charging is normally distributed with a mean ($\mu$) of 11 hours and a standard deviation ($\sigma$) of 2 hours.

(i) Write down the distribution of $Y$ by completing the following with the appropriate mathematical notation. (1 marks)

***Y ~ N (_____ , ______ )***

(ii) Find the probability that the phone could last less than 10 hours before it requires charging. (2 marks)

(iii) Find the probability that $Y$ is within 1 hour of its mean. (2 marks)

~~~ Please turn over for next question ~~~
~~~

**Question 4** (25 marks)

**Note**: For this question, a rare event is defined as the probability of an event happening being less than 5%.

(a) A truck driver is allowed to drive for a maximum of 10 hours and a maximum of 20 deliveries per day. His journey time per delivery is 30 minutes on average with a standard deviation of 8 minutes. Let *n* be the number of deliveries the driver is assigned to, and $\bar{T}$ be the mean journey time (in minutes) of the *n* deliveries.

(i) If $\bar{T}$ can be modelled by a normal distribution, complete the following with the appropriate mathematical notation. Cite one assumption and/or one theorem used, if any. (4 marks)

$\bar{T} \sim N$ (_____ , ______ )

(ii) What is the probability of the truck driver exceeding the maximum 10 hours driving limit, if n (the number of deliveries he is assigned to) is 16? (4 marks)

(iii) Is it a rare event for the truck driver to exceed the maximum 10 hours driving limit, if n (the number of deliveries he is assigned to) is 17? Justify your answer.

(4 marks)

(iv) Using your answer in part a (ii) and (iii) above, what is the maximum number of deliveries the truck driver should be assigned to each day in order to ensure that he has only a small chance of no more than 1 in 1000 of exceeding the maximum 10 hours driving limit? (2 marks)

(b) In a particular examination, the mean mark of all candidates in *City XYZ* was found to be 65, with a standard deviation of 10 marks. A random sample of 50 candidates from *City XYZ* was selected.

(i) Write down the distribution of $\bar{M}$ , the mean mark of the 50 students. Cite any theorem used (if any). (3 marks)

(ii) What is the probability that the mean mark of these 50 students exceeds 64?

(3 marks)

Another pack of 50 examination scripts with mean mark of 68.5 has surfaced.

(iii) What is the probability that the mean mark of 50 scripts exceeds 68.5, if the scripts were from *City XYZ*? (3 marks)

(iv) Based on your answer in part (iii) above, can we conclude that the 50 scripts were from *City XYZ*? Justify. (2 marks)

~~~ End of Paper ~~~
~~~