## GEA1000 QUANTITATIVE REASONING WITH DATA
## TUTORIAL 1

*Please work on the problems before coming to class. In class, you will engage in group work.*

1. The food delivery market has experienced tremendous growth, especially so during the Covid-19 pandemic. According to research firm Statista, revenue in the online delivery segment is estimated to be US$464 million in 2020[1].

Suppose you are a market researcher and would like to estimate the average income of delivery riders in Singapore, for June 2020.

(a) Discuss a suitable research question, population of interest, and sampling frame.

Following your discussion, your research team managed to obtain a list of contact numbers of all active delivery riders in the population of interest – the list contained a very large number of contacts. A simple random sample was drawn from the list, and 200 selected riders were surveyed regarding their incomes for the month of June 2020. The results are in the excel file "Food Delivery Data.xls".

For questions (b) to (e), please refer to the excel file "Food Delivery Data.xlsx".

*Data Details:*

| | |
|---|---|
| *Base* | *: Where the delivery rider is based – East or West of Singapore* |
| *Full/Part time* | *: Whether the delivery rider is working full or part time* |
| *Income* | *: The total income earned by the delivery rider for that month* |
| *Mode of Tpt* | *: '0' represents bicycles and motorcycles, '1' represents car and '2' represents walking* |

(b) Which of the above variables are numerical, and which are categorical? If the variables are categorical, are they ordinal or nominal?
(c) Calculate the average value of 'Mode of Tpt' for riders who are stationed in the East, and the West. How can we interpret these values?
(d) From the sample, what is the average income of all the riders (to 2 decimal places)?
(e) Your marketing team then published the results from (d). However, there were some skeptical full-time riders who feedbacked that the published average income seemed much lower than what they earn. What could be the issue?

CCNC

Source: https://www.straitstimes.com/singapore/manpower/new-players-ride-the-food-delivery-wave

2. Food delivery company ABC's riders use an app during their delivery process. The riders have recently feedbacked that the app was not balanced – sometimes in a short time period they are assigned too many deliveries to handle, at other times they have no nearby assignments at all. In response to those findings, the company developed a **new algorithm** for its riders' delivery app, hoping that it would be more balanced. The new algorithm has yet to be launched.

Suppose ABC now wants to know if the new algorithm is better than the old one and asks you to help plan an experiment over the course of one week. You are given authority and resources and asked to design an experiment on all 1206 riders in the company.

compare sd of orders per 4 hours

(a) Give a brief outline on how you would design the experiment.
*(Consider the 'other variables of concern', 'assignment' and 'blinding' issues, if any.)*

(b) We want to assign each rider to use an app that either utilises the new or old algorithm. However, due to logistical limitations, we can only afford to assign 500 riders to the new algorithm – the other 706 riders will be assigned the old algorithm. To conduct the assignment, the 1206 riders had their names placed on a list. The names were randomly shuffled, and 500 names were drawn. These drawn names were assigned to the new algorithm. The remaining 706 names were assigned to the old algorithm. Currently, the table below does not have all the information. How would you fill in the rest of the table?

| | Old Algorithm | New Algorithm | Total |
|---|---|---|---|
| Males | 464 | 360 | 824 |
| Females | 242 | 140 | 382 |
| Total | 706 | 500 | 1206 |

5076.02

3. When describing numerical variables in data sets, in addition to calculating the mean and standard deviation of these variables, it is a common practice for the description to include what is known as the “5-number summary” which consists of

   - Minimum 3950
   - Q1 4700
   - Median 5000
   - Q3 5500
   - Maximum 6300

   Recall that in the lecture videos, we have introduced the data set which gives information on the physical characteristics of 342 penguins across 3 different species. Use the data set penguins.csv together with a suitable software to give the 5 number summary for the mass of the Gentoo species of penguins. (We calculated the mean and standard deviation in the lecture videos)

4. Recall that in the lecture videos, we compared masses across different species of penguins and came across an *"observation".*

   ***Observation***
   - **Average mass** of Adelie and Chinstrap penguins were similar.
   However:
   - **Standard deviation** for Adelie penguins was larger than that for the Chinstrap.

   Some people may ask: 'Since the average mass is similar, shouldn't the spread of mass be similar too, between Adelie and Chinstrap?'. In the lecture videos, we stated some factors (gender, age, location) that could help us answer what could account for this difference in standard deviation. Your friend, Kowalski, suggests a possible explanation:

   *"Gender is the issue. The differences in mass between male and female penguins are greater for the Adelie species, compared to the Chinstrap species and this is the reason why the standard deviation for the Adelie species is higher than the Chinstrap despite having similar mean masses".*

   Describe how you would attempt to find out whether Kowalski's suggestion is valid. (Remember you cannot simply take individual penguins' data because the question is pertaining to the **species**.)

   In your description, include the following:
   - Formulation of a clear research question that you wish to investigate. (Hint: read the information above)
   - Measures of central tendencies that you would consider calculating along with any percentile calculations.
   - The extent to which you can answer the question using the "Penguins" data.

     You should clearly present your steps/calculations along with the rationale in your description.