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.

Research question: What is the average income of delivery riders in Singapore, for June 2020?
Type of research question: "Make an estimate about the population"
Population of interest: All delivery riders based in Singapore
Possible sampling frame: Contact numbers/email addresses of the delivery riders

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?

Numerical: Income. **Categorical Nominal**: Base, Full/Part time, Mode of Tpt

(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?

The average value of ‘Mode of Tpt’ among East riders = 0.663 (3 d.p.), while the average value of ‘Mode of Tpt’ among West riders = 0.667 (3 d.p.). We are unable to interpret these values meaningfully, as the data is **categorical nominal** in nature with three categories. Understanding the data you are working with is key to avoid misinterpretation of any data analyses performed.

(*Special case: if we had only two categories labelled ‘0’ and ‘1’ respectively, the average value could tell us specifically the proportion of individuals who belong to Category 1. For example, if we have 100 individuals, and the average value calculated is say, 0.49, we can infer that 49% of individuals belong to Category 1.)

(d) From the sample, what is the average income of all the riders (to 2 decimal places)?

Average income of all riders = $491.00 (rounded off 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?

Full-time riders' average income = $1465.11

Part-time riders' average income = $445.10

The majority of the sample consists of part-time riders (191 out of 200). Despite a random sample being obtained, nevertheless the response rate could have been low amongst the full-time riders (Possible explanation: the full-time riders might be busier than the part-time riders, so they would be less likely to respond to surveys) – more information is needed.

[1] 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.

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

As this is an experiment, we can try to **minimize the effects from other variables** that influence ("confound") the rating scores by the riders. For example, we can try to make all riders in the study rent standardized company cars/bikes so that the speed of delivery is consistent. Another example: we can try to ensure all riders are given the same working hours as well.

Random Assignment: This will make it very likely that characteristics of the subjects will be similar in both the treatment and control groups, especially for variables that the experiment is not able to manipulate – the riders' age, sex, driving experience, etc.

Double Blinding: It is important to blind both riders and assessors. Riders can just be given the app and not informed which algorithm they are using. The app should look and function the same regardless of the algorithm used. Since the new algorithm has not yet been launched, riders may not even know about its existence. In the same way, assessors can be blinded as well. Assessors need to be blinded because if they know which algorithm the rider is using, they may consider some of the rider's deliveries as valid/invalid depending on their internal bias.

(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?

From the given information, we know random assignment was conducted. However, it does not mean that the old and new algorithms must have the same number of people; nor does it mean that both algorithms must have the same number of males and females. Rather, we would expect to see both the old and new algorithm having a male:female ratio that is very close to (*note: not necessarily exactly*) 824 male and 382 female riders. One possible example is the following:

	Old Algorithm	New Algorithm	Total
Males	482	342	824
Females	224	158	382
Total	706	500	1206

(Rounded off to the nearest whole number)

Random assignment tends to make both groups (old and new algorithms in this case) similar in characteristics.

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
   - Q1
   - Median
   - Q3
   - Maximum

   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)

Using Excel, we can obtain the 5-number summary statistics as follows. (The minimum, maximum and median can be obtained using the Data Analytics Toolpak, whilst Q1 and Q3 can be obtained using the "=quartile" command.)

Minimum – 3950g

Q1 – 4700g

Median – 5000g

Q3 – 5500g

Maximum – 6300g

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.

We wish to investigate the following:

Question: Is gender the reason for the greater variation in mass among the Adelie species as compared to the Chinstrap species?

We will only use data from the Adelie and Chinstrap species and ignore the Gentoo species. For the Adelie penguins, some of the values under "Sex" take the value "NA". In this case, since they only make up a few data points out of more than 150 data points, we can omit them without compromising much on the accuracy and validity of our calculations for the data set.

We will next calculate the proportion of males and females between both species. The results are summarised below.

	Proportion of males	Proportion of females
Chinstrap	0.5	0.5
Adelie	0.5	0.5

We see that across both species, there is an equal proportion of males and females. This is informative because, in layman terms, it denies the claim that the Chinstrap species' lesser spread in mass is due to them having predominantly male or predominantly female penguins.

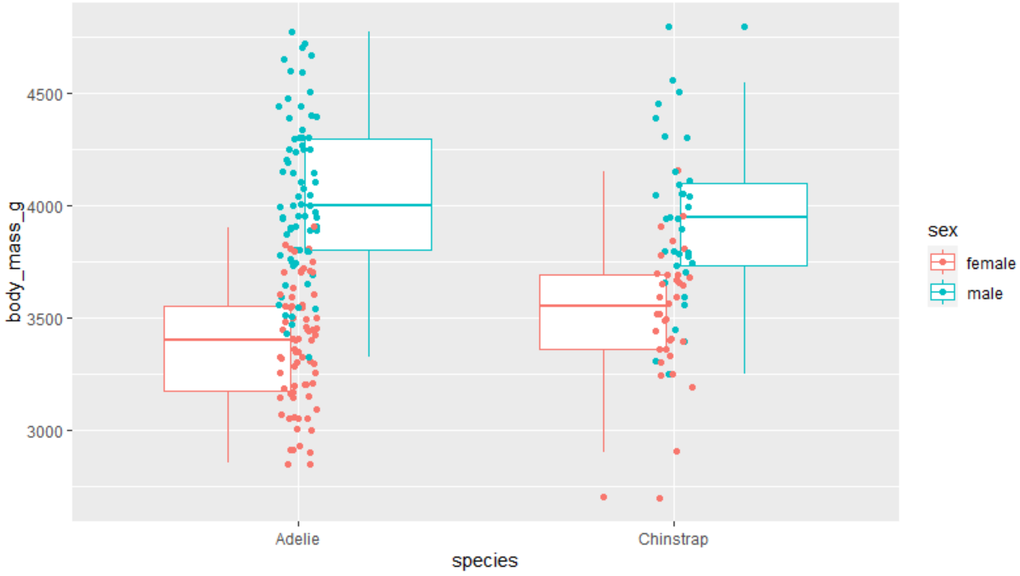
Next, let us look at the spread (in terms of SD) within males and females for each species.

	Standard deviation for males	Standard deviation for females	Overall standard deviation
Chinstrap	362.1	285.3	384.3
Adelie	346.8	269.4	458.6

Notice that the standard deviation for males across both species are more or less the same; likewise for females. One would expect the overall standard deviation to be the same as well, but note that this **does not take into account where the middle points of both species are.** So, it becomes worthwhile to investigate the measures of central tendency for each species.

	Mean mass	Median mass
Chinstrap Male	3938.97	3950
Chinstrap Female	3527.20	3550
Adelie Male	4043.49	4000
Adelie Female	3368.84	3400

From the table above, we see that for every subgroup, the mean and median are similar; hence, we can use boxplots as a visualisation tool, together with the data points. (We will discuss boxplots in more detail, including how to generate them, in subsequent chapters.)



The general idea we can get from the boxplot above is that if 2 sets of box plots are at approximately the same height with the same mid-points, then combining them will not result in a significant increase for the spread of the data. From the picture above, we see that the differences in middle points and boxplots between the males and females of the Adelie species are significantly greater than the differences for the Chinstrap species.

Therefore, it does seem to indicate that **gender** could be a factor that would explain the greater spread in the mass of the Adelie penguins. However, we should not be too hasty and immediately conclude that gender is the sole reason why the spread for the Adelie species is greater than the Chinstrap species.

Another factor we may want to consider is the **location**. For example, do penguins from colder locations have a heavier mass compared to those which come from warmer climates?

Notice that the data for the Chinstrap species is localized to one island (Dream Island) whereas the Adelie species is scattered across 3 islands (which includes Dream Island). To ensure fairness, we would want to compare whether the Adelie penguins from Dream Island still differ as much in their mass compared to the Chinstrap species.

Also, notice that **age** is not reflected in this data. For example, we cannot even be sure whether male Adelie penguins are heavier than females. It could well be that, within this data set, male Adelie penguins tend to be older while female Adelie penguins tend to be younger, and therefore, the difference in mass is due to age rather than gender. Therefore, whilst the data does offer a plausible reason, we would want to get information pertaining to other variables that could also influence the mass of the penguins, before concluding whether gender is **solely** responsible for the greater variation in mass for the Adelie species of penguin.