GEA1000 Quantitative Reasoning with Data
2021/2022 Semester 1
Midterm Test
Suggested Solutions

NOTE: There are two versions each of Questions 7 to 10.

Q1

The table below shows male and female patients undergoing two treatment types, X or Y. The outcome of the treatment is designated as either successful or unsuccessful. The success rates of the respective treatments across genders are also calculated.

	Male			**Female**		
	No. of Patients Treated	No. of Success	Rate of Success	No. of Patients Treated	No. of Success	Rate of Success
Treatment X	?	?	50%	40	32	80%
Treatment Y	?	?	?	?	?	60%
Total	100	50	50%	?	?	?

Unfortunately, some of the data is missing. We know that all missing values are non-zero. Which of the following statements is **necessarily true**?

(I) Simpson's Paradox is observed when the subgroups of Treatment X and Treatment Y are combined, when considering the relationship between gender and outcome.
(II) Treatment type is a confounder between the variables gender and outcome.

a) Only (I)

b) Only (II)

c) Both (I) and (II)

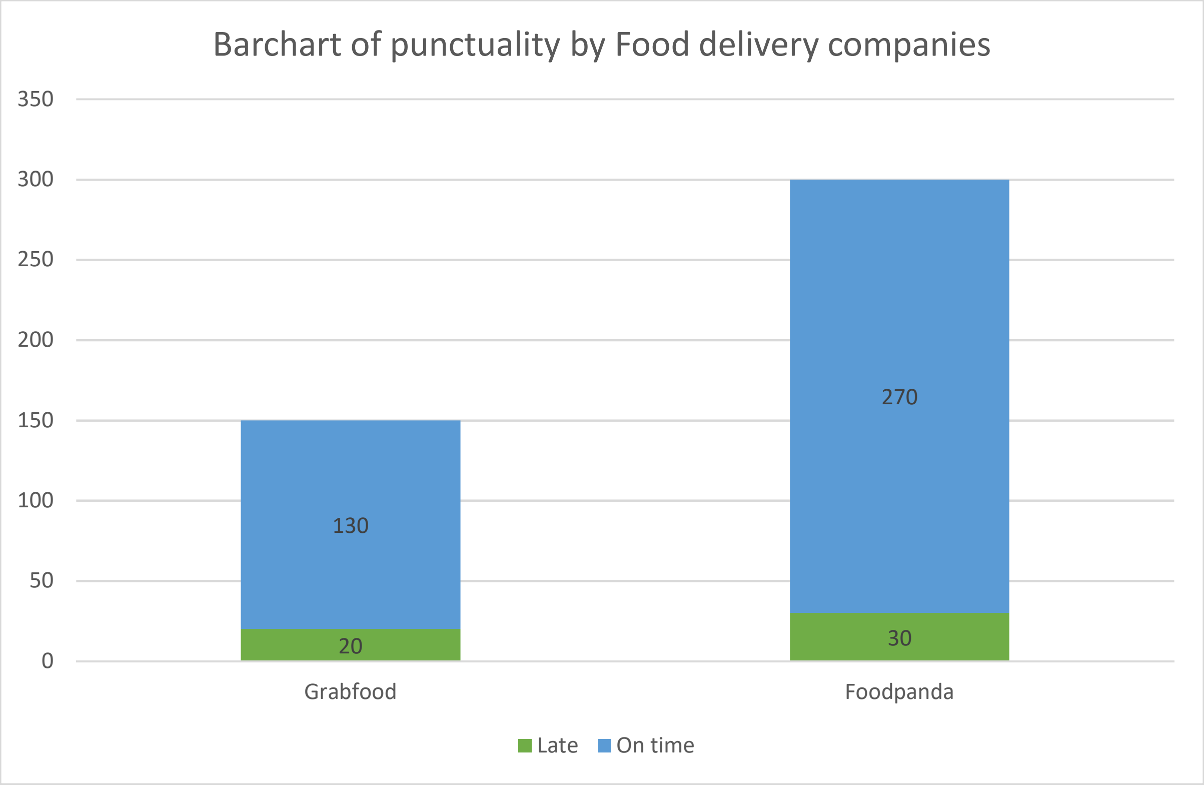
d) Neither (I) nor (II)

Explanation
By the basic rule of rates, among males, rate of success given Treatment Y must be 50%. By the basic rule of rates, among females, overall rate of success must be between 60% and 80%. Simpson's Paradox is **not** observed when the subgroups are combined, hence (I) is false.
Depending on the data, treatment type may or may not be a confounder.
For the following given set of values in the table, rate(X|male) = 80/100 = 40/50 = rate(X|female), so treatment type is not associated with gender. Treatment type is not necessarily a confounder hence (II) is false.

	Male			Female		
	Undergone Treatment	Successful	Rate of Success	Undergone Treatment	Successful	Rate of Success
Treatment X	80	40	50%	40	32	80%
Treatment Y	20	10	***50%***	10	6	60%
Total	100	50	50%	50	38	***>60% and <80%***

Q2

A group of market researchers were commissioned to investigate the relationship between two food delivery companies (Grabfood and Foodpanda) and their punctuality of deliveries (whether they are on time or late). The following chart displays a visual that the researchers used to aid in the presentation of their findings.



Which of the following statements is / are true based on the information given above? Select all options that apply.

a. Grabfood is positively associated with being on time for food deliveries.
b. Grabfood is positively associated with being late for food deliveries.
c. Foodpanda is positively associated with being on time for food deliveries.
d. Foodpanda is positively associated with being late for food deliveries.

Explanation

Let A be Grabfood. Let NA be Foodpanda.
Let B be Late. Let NB be On time.

Since rate(Late | Grabfood) = 20 / 150 > 30 / 300 = rate(Late | Foodpanda), there is a positive association between A and B. Hence, Grabfood is positively associated with being late for food deliveries. This also means that Foodpanda is negatively associated with being late for food deliveries, or that Foodpanda is positively associated with being on time for food deliveries.

Q3

A teacher has finished marking her students' test scripts for a Mathematics test. The maximum mark attainable for the test is 50. She records the following summary statistics for her class.

Mean : 32.4

Median : 31

Standard deviation : 13.32

Quartile 1 : 23

Quartile 3 : 38

Highest mark : 45

Lowest mark : 12

Range: 45 - 12 = 33

She returns the test papers to her class and goes through the answers. Whilst going through the answers, she realises that she has marked a question incorrectly for the whole class. She collects her students' scripts back and corrects her mistake. As a result, everyone in the class gets 2 additional marks. Which of the following summary statistics will change for the class?
Select all that apply.

a. Median
b. Standard deviation
c. Range
d. Quartile 3

Explanation
We have learnt in lecture that standard deviation does not change when a constant is added to all the data points and the median will change by the amount that is added to/subtracted from it. Quartile 3 is the 75th percentile and it will also change when constants are added to/subtracted from it. Range would not change since the maximum and minimum will change by the same amount.

Q4

A recent study revealed that Singapore is "the most tired country in the world, due to work and Internet". A researcher decided to conduct a further study on Internet usage behaviour and working hours among all Singaporean adults in Singapore. Data was collected by interviewing commuters alighting from Pasir Ris MRT (East), Woodlands MRT (North), Redhill MRT (South) and Jurong East MRT (West) from 8a.m. to 11p.m over a period of 7 days.

Which of the following statements is **necessarily true**?

a. As data was collected from different parts of Singapore, it is generalisable to the population of Singapore.
b. Due to the equal representation of Northern, Southern, Eastern and Western parts of Singapore, selection bias is minimised.
c. In this example, non-response bias exists because of a bad sampling plan.
d. None of the other options

Explanation
None of the options are correct: The selection of MRT Stations was not done by using a randomised mechanism. Therefore, data collected is not generalisable to the population of Singapore, and it is not a forgone conclusion that selection bias is minimised. There is a possibility that non-response bias exists in this example. However, the reason for its existence is not because of a bad sampling plan.

<u>Q5</u>

The drug ivermectin, which has all along been used for veterinary treatment of parasites, has come under the spotlight after calls for it to be included as part of Malaysia's list of treatments for COVID-19.

The focus on the drug came after news reports in April 2020, citing how researchers at Australia's Monash University had experimented with the drug and found that a single dose could stop the COVID-19 virus from growing in cell culture within 48 hours.

What are possible controls to compare the drug's effectiveness in treating COVID-19 symptoms, assuming that all drugs look and taste the same?
Select all that apply.

a. A drug made of glucose
b. Another drug that treats COVID-19 symptoms
c. A drug made of salt
d. An empty drug capsule

<u>Explanation</u>
A control group acts as a baseline for comparison with the treatment groups. The control group can simply be not receiving the treatment, a placebo or receiving an existing treatment instead.

Source: https://www.channelnewsasia.com/asia/what-ivermectin-and-why-malaysia-no-rush-approve-it-covid-19-treatment-2019206

<u>Q6</u>

The contingency table below shows the income level for a group of adults classified by gender.

		Income level			
		Low	Middle	High	Total
Gender	Male	23	30	29	82
	Female	26	28	20	74
	Total	49	58	49	156

The marginal rate, rate(Middle Income), is calculated to be _________ %; while the joint rate, rate(Female and Low Income), is calculated to be ________ %.
Give each answer as a percentage correct to 2 decimal places.

<u>Explanation</u>
To calculate the marginal rate, rate(Middle Income), we take the column total of all Middle Income persons divided by the grand total of everyone in the sample, ie. 58/156 = 37.18%. Then, to calculate the joint rate, rate(Female and Low Income), we take the count of "females with Low Income" divided by once again the grand total of everyone in the sample, ie. 26/156 = 16.67%.

Q7A

"A total of 600 COVID patients are in hospital, out of which 100 are seriously ill. Of those seriously ill patients, 50 of them require oxygen, and 75 of them are above 60 years old. 200 of these COVID patients require oxygen."

From the information contained in the above quote, which of the following is **necessarily true** regarding patients in this hospital?

(A) 200/600 of COVID patients who are not seriously ill require oxygen
(B) There is a positive association between requiring oxygen and being seriously ill.
(C) There is a positive association between being above 60 years old and being seriously ill.

Explanation
(A) is false because it should be 150/500 = 0.3
(B) is true because rate(requiring oxygen| not seriously ill) = 150/500 < 50/100 = rate(requiring oxygen| seriously ill)
(C) is false because rate(above 60 years old| not seriously ill) is unknown

Q7B
"A total of 600 COVID patients are in hospital, out of which 100 are seriously ill. Of those seriously ill patients, 75 of them require oxygen, and 50 of them are above 60 years old. 200 of these COVID patients are above 60 years old."

From the information contained in the above quote, which of the following is **necessarily true** regarding patients in this hospital?

(A) 200/600 of COVID patients who are not seriously ill are above 60 years old
(B) There is a positive association between being seriously ill and requiring oxygen.
(C) There is a positive association between being seriously ill and being above 60 years old.

Explanation
(A) is false because it should be 150/500 = 0.3
(B) is false because rate(oxygen required| not seriously ill) is unknown
(C) is true because
rate(above 60 years old| not seriously ill) = 150/500 < 50/100 = rate(above 60 years old| seriously ill)

Q8A

For the year 2020, the marginal death rate of Country A is greater than the death rate among the females of Country A, or in other words, rate(death) > rate(death | female). Which of the following statements must be true in Country A for the year 2020?

(A) rate(male) < rate(male | death)
(B) rate(male) > rate(male | death)
(C) rate(male) = rate(male | death)

Explanation

By the basic rule of rates, rate(death | male) > rate(death) > rate(death | female), so symmetry rule gives us rate(male | death) > rate(male | no death). By the basic rule of rates again, rate(male | death) > rate(male) > rate(male | no death).

Q8B

For the year 2020, the marginal death rate of Country A is greater than the death rate among the females of Country A, or in other words, rate(death) > rate(death | female). Which of the following statements must be true in Country A for the year 2020?

(A) rate(female) < rate(female | death)
(B) rate(female) > rate(female | death)
(C) rate(female) = rate(female | death)

Explanation

By the basic rule of rates, rate(death | male) > rate(death) > rate(death | female), so symmetry rule gives us rate(female | no death) > rate(female | death). By the basic rule of rates again, rate(female | no death) > rate(female) > rate(female | death).

Q9A

A researcher wanted to test if a new drug X was effective in reducing headaches. The study contained 1000 subjects with the following characteristics:

	Young	Old	Row Total
Male	150	150	300
Female	350	350	700
Column Total	500	500	1000

Random assignment was conducted to assign the 1000 subjects into the treatment and the control groups. 480 subjects were assigned to the treatment group, while the remaining 520 subjects were assigned to the control group.

How many young males will we likely see in the treatment group?

a. About 36
b. About 72
c. About 144
d. About 150

Explanation

There are 150 young males in the study. If the random assignment ratio is 480:520 for the treatment and control group respectively, it is likely to see about 150*480/1000 = 72 young males in the treatment group.

Q9B

A researcher wanted to test if a new drug Z was effective in curing stomachaches. The study contained 2000 subjects with the following characteristics:

	Young	Old	Row Total
Male	600	600	1200
Female	400	400	800
Column Total	1000	1000	2000

Random assignment was conducted to assign the 2000 subjects into the treatment and the control groups. 980 subjects were assigned to the treatment group, while the remaining 1020 subjects were assigned to the control group.

How many young males will we likely see in the treatment group?

a. About 147
b. About 294
c. About 588
d. About 600

Explanation

There are 600 young males in the study. If the random assignment ratio is 980:1020 for the treatment and control group respectively, it is likely to see about 600*980/2000 = 294 young males in the treatment group.

Q10A

In a research study investigating the association between fried food consumption and mortality, data was collected on the consumption of fried food, mortality, and possible confounding variables. Participants were asked to report on their fried food consumption (none, <1 serving per week, 1-2 servings per week, 3-6 servings per week, or ≥ 1 serving per day), race/ethnicity (White, African-American, Hispanic, or others), and annual income (<20 000, 20 000-49 999, or > 50 000 dollars). Weight and height at baseline were measured during clinic visits using standard methods. Body mass index was calculated as weight (in kilograms) divided by height (in meters) squared.

What types of variables are fried food consumption, race, annual income, and body mass index, respectively?

Answer
Ordinal, nominal, ordinal, numerical

Q10B

In a research study investigating the association between consumption of processed food and mortality, data was collected on processed food consumption, mortality, and possible confounding variables. Weight and height at baseline were measured during clinic visits using standard methods. Participants were asked to report on their consumption of processed food (none, < 2 servings per month, 2-3 servings per month, or ≥ 1 serving per week), smoking status (never, former, current) , and race/ethnicity (Chinese, Malay, Indian, or others).

What types of variables are weight, processed food consumption, smoking status, and ethnicity, respectively?

Answer
Numerical, ordinal, nominal, nominal