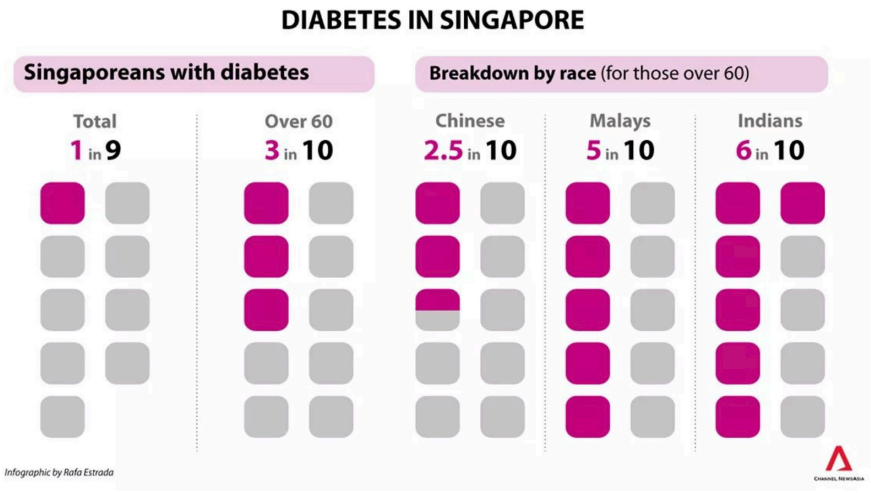
Exercise 2

1. The figure below shows that out of every 9 Singaporeans, 1 of them has diabetes. Similarly, out of 10 Singaporeans over 60, 3 of them have diabetes. Let us define "over 60" as old and "60 and below" as young. Which of the following statements is/are true? Select all that apply.



(A) rate(Diabetes | Young) > rate(Diabetes | Old).

(B) rate(Young | Diabetes) < rate(Young | No diabetes).

(C) rate(Old | Diabetes) < rate(Old | No diabetes).

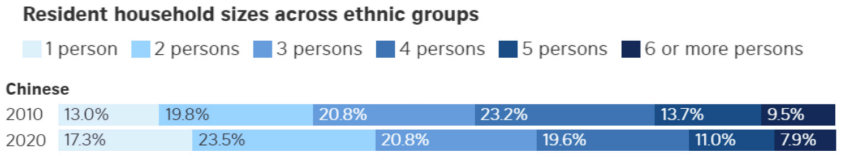
(D) rate(Diabetes | Young) < rate(Diabetes | Old).

(B) and (D) are correct. From the above, we can see that the overall rate of diabetes is $1/9 = 0.111$. Additionally, we also know that rate(Diabetes | Old) = 0.3. Using the basic rule of rates, we can deduce that rate(Diabetes | Young) must be less than 0.111, and by extension, less that rate(Diabetes | Old).

Since rate(Diabetes | Young) < rate(Diabetes | Old), it must also mean that rate(Young | Diabetes) < rate(Young | No diabetes).

Since rate(Diabetes | Old) > rate(Diabetes | Young), we can also conclude that rate(Old | Diabetes) > rate(Old | No diabetes).

2. On 19 June 2021, The Straits Times published the figure below, taken from a population census of Singapore.



Use only the information shown in the figure to answer the following question.

Suppose that households with 1-3 people are considered "small" whereas those with 4 or more people are considered "large". Which of the following statements is/are true among Chinese resident households in the years 2010 and 2020? Select all that apply.

(A) The year 2020 is positively associated with small households.

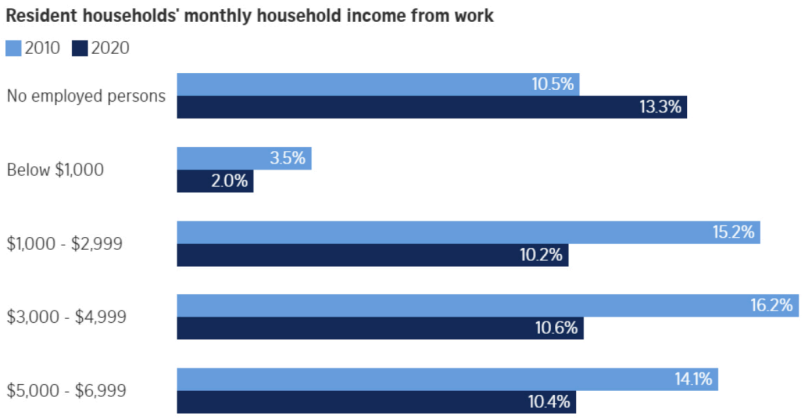
(B) The year 2020 is positively associated with large households.

(C) The year 2010 is positively associated with small households.

(D) The year 2010 is positively associated with large households.

(A) and (D) are correct. Note that rate(Small | 2010) = 53.6%, rate(Small | 2020) = 61.6%. Since rate(Small | 2020) > rate(Small | 2010), this means that the year 2020 is positively associated with small households. This is the same as saying that in 2010, there are more large households, thus there is also positive association between the year 2010 and large households.

3. On 19 June 2021, The Straits Times published the figure below, taken from a population census of Singapore. Each household may only belong to a single category.



What can be said about the resident households, earning more than $6,999 from work? From the following statements, select all that apply.

(A) A majority of resident households are earning more than $6,999 from work in 2020.

(B) A larger proportion of resident households are earning more than $6,999 from work in 2020, as compared to 2010.

(C) rate(Income > $6,999 | 2020) > rate(Income > $6,999 | 2010). Here "Income" represents Household monthly income from work.

(D) rate(Income > $6,999 | 2020) < rate(Income > $6,999 | 2010). Here "Income" represents Household monthly income from work.

(A), (B) and (C) are correct. Proportions must add up to 100%. In 2020, the proportion earning more than $6,999 is

$$100\% - (13.3 + 2 + 10.2 + 10.6 + 10.4)\% = 53.5\%.$$

Thus a majority of resident households are earning more than $6,999 in 2020.

In 2010, the proportion earning more than $6,999 is

$$100\% - (10.5 + 3.5 + 15.2 + 16.2 + 14.1)\% = 40.5\%.$$

This is smaller than in 2020.

Recall that proportions are the same as rates. Since 53.5% is greater than 40.5%, this means that rate(Income > $6,999 | 2020) is greater than rate(Income > $6,999 | 2010), showing that a larger proportion is earning more than $6,999 in 2020, as compared to 2010.

4. How does "forgiveness" (being forgiving) and empathy go together? The study of Toussaint and Webb on 45 men and 82 women are summarised in the following hypothetical tables:

Distribution of 45 men

	Empathy	No empathy	Row total
Forgiving	10	10	20
Not forgiving	9	16	25
Column total	19	26	45

Distribution of 82 women

	Empathy	No empathy	Row total
Forgiving	30	31	61
Not forgiving	12	9	21
Column total	42	40	82

Which of the following statements is/are true?

(I) Forgiveness and empathy are positively associated among men.

(II) Forgiveness and empathy are positively associated among women.

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

Answer is (A). Among men,

$$\text{rate(Empathy} \mid \text{Forgiving)} = \tfrac{10}{20} = 0.5,$$

$$\text{rate(Empathy} \mid \text{Not forgiving)} = \tfrac{9}{25} = 0.36.$$

So there is positive association between forgiveness and empathy among men. Among women, the corresponding rates are 0.49 and 0.57, so forgiveness and empathy are negatively associated.

5. The contingency table below shows the classification of hair descriptions of students studying in an international school in Singapore.

	Hair type				
	Straight		Curly		
Hair colour	Male	Female	Male	Female	**Total**
Red	7	9	8	5	29
Brown	35	20	12	16	83
Blonde	51	55	38	27	171
Black	22	25	19	24	90
Total	115	109	77	72	373

The marginal rate, rate(Curly), is _____%; while the joint rate, rate(non-Black and Female) is _____%. Give each answer as a percentage correct to 2 decimal places.

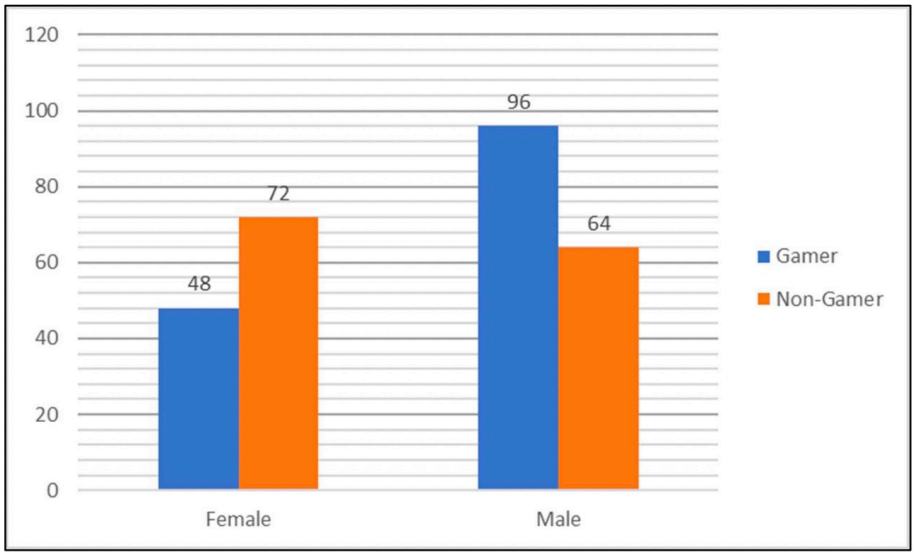
To calculate the marginal rate, rate(Curly), we take the column totals of all Curly-haired persons (both Male and Female) dvided by the grand total of everyone in the data set, i.e.

$$\frac{(77+72)}{373} \approx 39.95\% \text{ (2 decimal places).}$$

Then, to calculate the joint rate, rate(non-Black and Female), we take the count of "Females with non-black hair" divided by once again the grand total of everyone in the data set, i.e.

$$\frac{(9+20+55+5+16+27)}{373} \approx 35.39\% \text{ (2 decimal places).}$$

6. The bar graph below shows the number of gamers and non-gamers among males and females. Which of the following statements is/are true?



(A) There is a negative association between being female and being a gamer since rate(Female | Gamer) = 0.33 is less than rate(Female | Non-Gamer) = 0.53.

(B) There is a negative association between being female and being a gamer since rate(Gamer | Female) = 0.4 is less than rate(Gamer | Male) = 0.67.

(C) There is a negative association between being female and being a gamer since rate(Female | Gamer) = 0.33 is less than rate(Male | Gamer) = 0.67.

Answer is (A). To establish association, we should be comparing rate(A | B) and rate(A | NB), thus the comparison of rate(Female | Gamer) and rate(Male | Gamer) is irrelevant. Based on the graph, the contingency table can be constructed as shown:

	Female	Male	Row total
Gamer	48	96	144
Non-Gamer	72	64	136
Column total	120	160	280

Since rate(Gamer | Male) $= \frac{96}{160} = 0.6$, it is incorrect to say that rate(Gamer | Male) = 0.67. Lastly,

$$\text{rate(Female | Gamer)} = \frac{48}{144} = 0.33,$$

$$\text{rate(Female | Non-Gamer)} = \frac{72}{136} = 0.53.$$

Since rate(Female | Gamer) < rate(Female | Non-Gamer), there is negative association between being female and being a gamer.

7. There are (another) two types (A and B) of possible treatment for kidney stones. The number of patients given each treatment and the number of successful outcomes, according to treatment type and stone size are shown in the table below.

	Treatment A		Treatment B	
Stone size	Number	Success	Number	Success
Small	87	81	270	234
Large	263	192	80	55
Total	350	273	350	289

Which of the following statements is/are correct? Select all that apply.

(A) Treatment A has a higher success rate than treatment B in each of the two groups (small stone size and large stone size).

(B) When groups of patients with small stones and large stones are combined, treatment B has a higher success rate than treatment A.

(C) There is positive association between treatment A and success among patients with small stones, and also among patients with large stones.

(D) When groups of patients with small stones and large stones are combined, treatment A and success have a negative association.

(E) There is no association between stone size and success, as rate(Sucess | Small stones) is equal to rate(Success | Large stones).

(A), (B), (C) and (D) are correct. For patients with small stones, the rate of success for treatment A is $\frac{81}{87} = 93\%$ while the rate of success for treatment B is $\frac{234}{270} = 87\%$. For patients with large stones, the rate of success for treatment A is $\frac{192}{263} = 73\%$ while the rate of success for treatment B is $\frac{55}{80} = 69\%$. So treatment A has a higher success rate than treatment B in each of the two groups.

When the two groups are combined, the rate of success for treatment A is $\frac{273}{350} = 78\%$ while the rate of success for treatment B is $\frac{289}{350} = 83\%$. Thus it is true that when groups of patients with small stones and large stones are combined, treatment B has a higher success rate than treatment A. In other words, since rate(Success | Treatment A) is smaller than rate(Success | Treatment B), treatment A and success have a negative association.

There is positive association between treatment A and success among patients with small stones, since rate(Success | Treatment A)= 93% which is larger than rate(Success | Treatment B)= 87%. Similarly, there is positive association between treatment A and success among patients with large stones, since rate(Success | Treatment A)= 73% which is larger than rate(Success | Treatment B)= 69%.

It is incorrect to say that there is no association between stone size and success. We see that

$$\text{rate(Success} \mid \text{Small)} = \frac{315}{357} = 0.88;$$

which is larger than

$$\text{rate(Success} \mid \text{Large)} = \frac{247}{343} = 0.72.$$

8. Joseph conducted a study on night owls (individuals who sleep after 12am on average every night) among staff and students in NUS. He gathered the following result on rates:

$$\text{rate(Night owl} \mid \text{Student)} = 0.7;$$

$$\text{rate(Non-Night owl} \mid \text{Staff)} = 0.4.$$

Which of the following statements is correct?

(A) rate(Night owl) must be between 0.4 and 0.7.

(B) rate(Night owl) must be between 0.6 and 0.7.

(C) rate(Non-Night owl) must be between 0.4 and 0.7.

(D) rate(Non-Night owl) must be between 0.6 and 0.7.

Answer is (B). From rate(Non-Night owl | Staff)= 0.4, we obtain

$$\text{rate(Night owl} \mid \text{Staff)} = 1 - 0.4 = 0.6.$$

Since rate(Night owl | Student) = 0.7, applying the basic rule of rates, the overall rate(Night owl) must be between these two rates at the subgroups. Thus rate(Night owl) must be between 0.6 and 0.7. In addition, from rate(Night owl | Student)= 0.7, we obtain

$$\text{rate(Non-Night owl | Student)} = 1 - 0.7 = 0.3.$$

Applying the basic rule of rate, the overall rate(Non-Night owl) must be between 0.3 and 0.4.

9. A newspaper article had a headline "30% of local university students admitted last year graduated from a polytechnic". Assume there are only 2 universities (Uni A and Uni B). In Uni A, 50% of its local students admitted last year graduated from a polytechnic. In Uni B, the percentage of its local students admitted last year who graduated from a polytechnic must be

   (A) 10%.

   (B) 40%.

   (C) between 30% and 50%.

   (D) less than 30%.

   (E) more than 50%.

   Answer is (D). By the basic rule on rates, since the overall rate is 30% and the rate at Uni A is 50%, the rate at Uni B must be less than 30%.

10. The relative frequency table below shows the distribution of annual total personal income for the entire population of 6,402,386 living in City X. It is known that there are 59% males and 41% females in City X.

Income	Percent
\$9,999 or less	2.2%
\$10,000 to \$14,999	4.7%
\$15,000 to \$24,999	15.8%
\$25,000 to \$34,999	18.3%
\$35,000 to \$49,999	21.2%
\$50,000 to \$64,999	13.9%
\$65,000 to \$74,999	5.8%
\$75,000 to \$99,999	8.4%
\$100,000 or more	9.7%

   We are told that in City X, 71.8% of females earn less than \$50,000 per year. Which of the following statements is correct?

   (A) There is positive association between being male and earning less than \$50,000.

   (B) There is negative association between being male and earning less than \$50,000.

   (C) There is no association between being male and earning less than \$50,000.

   (D) We do not have sufficient information to determine the correctness of the other three statements.

   Answer is (B). We need to figure out the relationship between rate(Less than 50,000 | Males) and rate(Less than 50,000 | Females). The latter is given in the question as 71.8%. From the table, one can calculate the overall rate(Less than 50,000) to be $2.2+4.7+15.8+18.3+21.2 = 62.2\%$. So among everyone, the rate of those earning less than 50,000 is 62.2%, while for females it is higher at 71.8%. By basic rule on rates, the rate among males should be less than 62.2% and hence is less than 71.8%. Thus

   $$\text{rate(Less than 50,000 | Males)} < \text{rate(Less than 50,000 | Females)}$$

   which implies that the association between male and earning less than 50,000 is negative.

11. The website correlated.org presents the following for December 24th 2018. 352 people are surveyed, of whom 131 find the sound of windshield wipers to be soothing. Among the 352 people, 55% stay in the movie theater until the credits end. But among those who find the sound of windshield wipers to be soothing, 75% stay in the movie theater until the credits end. Among those who do not find the sound of windshield wipers to be soothing, what would be the percentage who stay in the movie theater until the credits end?

    (A) More than 75%.

    (B) Equal to 75%.

    (C) More than 55% and less than 75%.

    (D) Equal to 55%.

    (E) Less than 55%.

    Answer is (E). Note that 55% is the overall rate and 75% is the rate in the group who find the sound of windshield wipers soothing. By the Basic Rule on Rates, 55% must be between 75% and the rate in the other group, which must be less than 55%.

12. By "elderly", we mean a person who is more than 65 years old. In Singapore, the percentage of elderlies among women is higher than the percentage of elderlies among men. Which of the following statements is/are true?

    (I) In Singapore, the percentage of women among elderlies is higher than the percentage of women among the non-elderlies.

    (II) In Singapore, the percentage of women is higher than the percentage of men among elderlies.

    (A) Only (I).

    (B) Only (II).

    (C) Both (I) and (II).

    (D) Neither (I) nor (II).

    Answer is (A). From the information given,

    $$\text{rate(Elderlies} \mid \text{Women)} > \text{rate(Elderlies} \mid \text{Men)}$$

    and thus women and elderlies are positively associated. This also means that

    $$\text{rate(Women} \mid \text{Elderlies)} > \text{rate(Women} \mid \text{Non-elderlies),}$$

    which is expressed by statement (I). We cannot determine the percentage of women and men among the elderlies with the information given.

13. Consider the following statements.

    (I) A spokesman was quoted as saying that the proportion of a certain ethnic group among those who contracted Disease X was lower than the proportion of that ethnic group in the general population.

    (II) A reporter interpreted the statement and concluded that the members of this ethnic group are less likely to contract Disease X than a random member of the population.

    Which of the following is correct?

    (A) The two statements are equivalent.

    (B) We can infer statement (I) from (II) but not the other way round.

    (C) We can infer statement (II) from (I) but not the other way round.

(D) We can neither infer statement (I) from (II), nor infer statement (II) from (I).

Answer is (A). Let M represent the particular ethnic group (NM represent not the ethnic group), C represent contracting Disease X (NC represent not contracting Disease X). Statement (I) says that

$$\text{rate}(M \mid C) < \text{rate}(M).$$

By the basic rule on rates,

$$\text{rate}(M \mid C) < \text{rate}(M) < \text{rate}(M \mid NC).$$

This implies

$$\text{rate}(C \mid M) < \text{rate}(C) < \text{rate}(C \mid NM).$$

In particular,

$$\text{rate}(C \mid M) < \text{rate}(C),$$

which is statement (II).

14. Su is investigating the association between blood pressure and "workaholism" in a certain population. Someone who works more than 75 hours per week is considered a workaholic.

    The income level and blood pressure (high or normal) for each subject and whether or not they are classified as "workaholic" are recorded and summarised in the table below. Here "HBP" denotes "high blood pressure" while "NBP" denotes "normal blood pressure".

	Income Group					
	Low		Middle		High	
	HBP	NBP	HBP	NBP	HBP	NBP
Workaholic	25	75	23	87	26	134
Non-workaholic	25	80	18	72	9	51

    Which of the following statements is true?

    (A) For subjects in the "Middle" income level group, there is a positive association between being a "workaholic" and having "high blood pressure".

    (B) For subjects in the "Middle" income level group, there is no association between being a ''workaholic" and having "high blood pressure".

    (C) For subjects in the "Middle" income level group, there is a negative association between being a "workaholic" and having "high blood pressure".

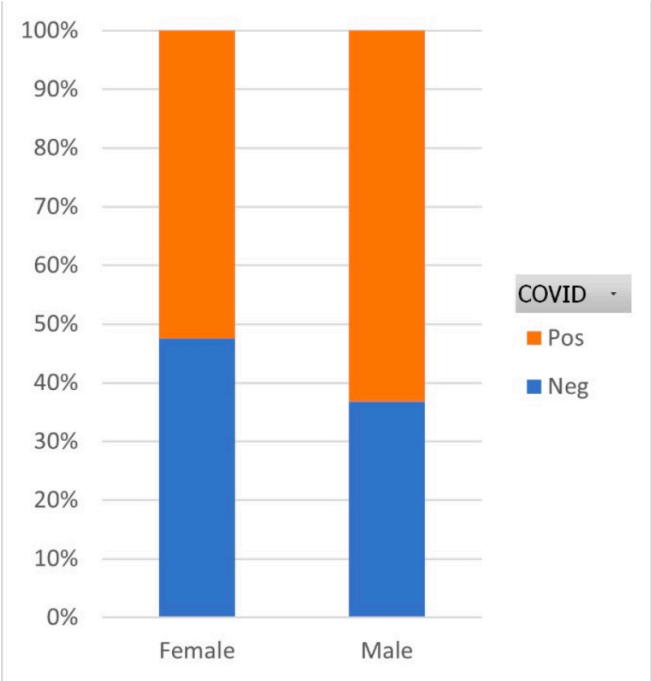
    Answer is (A). There is positive association between being a "workaholic" and having "high blood pressure" since (in the "Middle" income level group)

$$\text{rate}(\text{HBP} \mid \text{Workaholic}) = \frac{23}{(23+87)} = 0.21,$$

    which is larger than

$$\text{rate}(\text{HBP} \mid \text{Non-workaholic}) = \frac{18}{(18+72)} = 0.2.$$

15. The graph below shows the stacked bar plot for the rate of COVID infection among males and females in Country X. Which of the folowing variables must be positively associated with each other? Select all that are true.



(A) Female and COVID-positive.
(B) Male and COVID-positive.
(C) Female and COVID-negative.
(D) Male and COVID-negative.

(B) and (C) are correct. Based on the graph,

$$\text{rate(Positive} \mid \text{Male)} > \text{rate(Positive} \mid \text{Female)}.$$

So, males and COVID-positive must be positively associated. Similarly,

$$\text{rate(Negative} \mid \text{Female)} > \text{rate(Negative} \mid \text{Male)}.$$

So, females and COVID-negative must be positively associated.

16. The Lord of the Rings: The Fellowship of the Ring was released in December 2001. Suppose that

(I) Among the people in Singapore who were born before 2000, 10% watched the film.
(II) Among the people in Singapore who were born during or after 2000, 20% watched the film.

Choose the best option below. Among all the people in Singapore, the percentage who watched the film _____.

(A) must be 15%.
(B) must be between 10% and 20%.
(C) can be less than 10%.
(D) can be more than 20%.

Answer is (B). By the basic rule on rates, since 10% and 20% are the respective rates in the two groups, the overall rate must be between 10% and 20%.

17. Coriander is a common herb used to add flavour to various kinds of dishes. Suppose we have two countries: Country X and Country Y. We have individuals who either dislike coriander, or like coriander, and are either male or female. We know that in country X,

- rate(Dislike) = 0.1.
- rate(Dislike | Male) = 0.3.

Also, in country Y,

- rate(Female) = 0.8.
- rate(Female | Dislike) = 0.4.

Which of the following statements must be true? Select all that apply.

(A) rate(Male) < rate(Female) in country X.

(B) There is a positive association between disliking coriander and females in country X and country Y separately.

(C) Overall rate(Male) is between 0.2 and 0.5, when both countries are combined.

Only (A) is correct. The Basic Rule on Rates states that

$$\text{rate(A | not B)} \leq \text{rate(A)} \leq \text{rate(A | B)}.$$

As such, in country X, we note that rate(Dislike | Female) has to be between 0 and 0.1, since rate(Dislike) = 0.1 and rate(Dislike | Male)= 0.3. In any case, rate(Dislike) is closer to rate(Dislike | Female) than it is to rate(Dislike | Male). Hence, rate(Male) < rate(Female) in country X is correct.

In country X,

$$\text{rate(Dislike | Male)} > \text{rate(Dislike | Female)}.$$

Therefore, there is a negative association between disliking coriander and being females in country X. In a similar way, using the Basic Rule on Rates, we note that in country Y,

$$\text{rate(Female | Dislike)} < \text{rate(Female)} < \text{rate(Female | Not Dislike)}.$$

Therefore, there is a negative association between disliking coriander and being females and statement (B) is incorrect.

Finally, in country Y,

$$\text{rate(Male)} = 1- \text{rate(Female)} = 0.2.$$

On the other hand, we note previously in country X that females should form the majority as rate(Female) > rate(Male). Hence, rate(Male) can be smaller than 0.2. Thus the overall rate(Male) can be between 0 and 0.2 and statement (C) is incorrect.

18. Which of the following statements is **true**?

(A) Confounders will always lead to Simpson's Paradox.

(B) An observational study can be conducted when the researcher is unable to assign participants into the treatment and control groups.

(C) Randomised assignment will always result in equal allocation of the number of subjects across the treatment and control groups.

(D) Observational studies are better at showing causation than experimental studies because random assignment is used in observational studies to minimise the effect of confounding variables.

Answer is (B). By definition, an observational study occurs when the researcher does not get to decide if the participant receives the treatment or not. Simpson's paradox is always due to a confounder, but a confounder does not always lead to Simpson's paradox being observed. Random assignment tends to lead to two groups that exhibit similar characteristics, if there is large enough number of subjects, but not necessary equal number of subjects in both groups. In observational studies, researchers are unable to assign the participants into control and treatment groups.

19. The rate of lung cancer among females in Singapore is 40%, while the rate of lung cancer among males in Singapore is also 40%. Researchers also discovered that the rate of lung cancer among smokers in Singapore is 70%. Which of the following statements is/are true?

    (I) Sex is a confounder when discussing the relationship between smoking and lung cancer.

    (II) Lung cancer is positively associated with smoking.

    (A) Only (I).

    (B) Only (II).

    (C) Neither (I) nor (II).

    (D) Both (I) and (II).

    Answer is (B). We know that rate(Lung cancer | Females) is 40% which is also equal to rate(Lung cancer | Males). Hence, lung cancer and being a male/female are not associated, which means that sex cannot be a confounder when looking at the relationship between smoking and lung cancer.

    By the basic rule on rates, we know that rate(Lung cancer) = 40%. Given that rate(Lung cancer | Smokers) = 70%, we know that rate(Lung cancer | Non-smokers) must be less than 40%. Since rate(Lung cancer | Smokers) is greater than rate(Lung cancer | Non-smokers), we know that lung cancer is positively associated with smoking.

20. A researcher wants to find out if drinking tea helps to reduce memory loss. He interviewed 100 elderly citizens from an Elder Care Center and inquired if they were tea drinkers. 60 of them were classified as tea drinkers, while the remaining 40 were not. He then asked them to play a specific memory game to test their memory. The researcher also noted that a potential confounding variable was "gender". To control for this potential confounder (gender), the researcher could perform

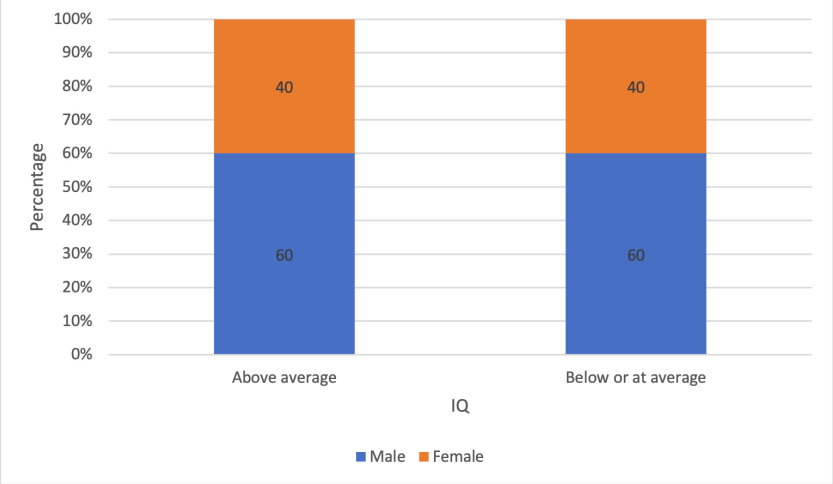
    (A) double blinding.

    (B) random assignment.

    (C) slicing of the data.

    Answer is (C). This is an observational study, and thus only slicing can be done to control for this potential confounder. Random assignment is only suitable for experimental studies. Double blinding also does not control for the effects of gender as a confounder.

21. A researcher is interested in finding out if gender affects whether a student is left-handed. The information on gender (Male/Female), master hand (Left/Right) and IQ (Above average/Below or At average) of all students was collected. The data was used to plot the figure below.



The researcher makes the following two statements.

(I) IQ can be a confounder when investigating the relationship between gender and master hand.

(II) When finding out if IQ affects whether a person is left-handed, it is possible that Simpson's Paradox is observed in the data due to the third variable, gender.

Which of the following correctly describes the two statements above?

(A) Statement (I) is true but statement (II) is false.

(B) Statement (I) is false but statement (II) is true.

(C) Both statements are true.

(D) Both statements are false.

Answer is (D). From the graph, we see that

$$\text{rate(Male | Above average IQ)} = \text{rate(Male | Below or at average IQ)} = 0.6.$$

Alternatively, from the graph, we see that the ratio of male to female in the "Above average" category is the same as in the "Below or at average" category. Hence, IQ is not associated with gender. Since IQ is not associated with gender, it cannot be a confounder when investigating the relationship between gender and master hand. Thus statement (I) is false.

Since gender is not associated with IQ, it cannot be a confounder when studying the relationship between IQ and master-hand. Since it is not even a confounder in the first place, Simpson's Paradox cannot be observed. Thus, statement (II) is false.

22. A researcher conducted an observational study to investigate whether smoking is associated with heart disease. 1000 participants were recruited in the study and the researcher obtained the following result:

	Heart disease	No heart disease
Smoker	146	317
Non-smoker	324	213

He also observed that 80% of the smokers were alcoholics, while 85% of the people with heart disease were alcoholics. Consider the following statements:

(I) There is positive association between smoking and heart disease.

(II) Being an alcoholic is a confounder.

Based only on the information given above, which of the two statements must be true?

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

Answer is (C). To conclude if being an alcoholic is a confounder, we need information on the percentage of non-smokers who are alcoholic and the percentage of people who are alcoholics among those who do not have heart disease.

From the table,

$$\text{rate(Smokers | Heart disease)} = \frac{146}{470} = 0.31$$

and

$$\text{rate(Smokers | No heart disease)} = \frac{317}{317+213} = 0.60.$$

Thus, in this example, smoking is negatively associated with heart disease, since the rate of smokers is lower in the heart disease group as compared to the group with no heart disease.

23. Consider a study that intends to examine whether the colour red makes children act impulsively. A group of 500 children were assigned into two groups by the expert opinion of a child psychologist; group Red if the psychologist pointed to the child, and group Green if the psychologist did not. Each child is then led into a room that has a big button in the colour of their group and labelled "DO NOT PRESS ME!". It is then recorded whether the child presses the button within 10 minutes. All the children were each then given a candy for participating.

    Which of the following conclusions from the analysis of data can establish that wearing spectacles (whether the child wears spectacles) is a confounder in this study?

    (A) Wearing spectacles is positively associated with being in group Red, and negatively associated with pressing the button.

    (B) Wearing spectacles is associated with being in group Red, and is not associated with pressing the button.

    (C) Wearing spectacles is not associated with being in group Red, and is not associated with pressing the button.

    (D) None of the other given options is correct.

    Answer is (A). A confounder is a third factor that is associated with both the exposure and response variables. The directions of the associations do not matter.

24. Consider a study that intends to examine whether the colour red makes children act impulsively. A group of 500 children were assigned into two groups by the expert opinion of a child psychologist; group Red if the psychologist pointed to the child, and group Green if the psychologist did not. Each child is then led into a room that has a big button in the colour of their group and labelled "DO NOT PRESS ME!". It is then recorded whether the child presses the button within 10 minutes. All the children were each then given a candy for participating.

    The children were also asked if they like candies. The following table summarises the data. For instance, 22 children from group Red that pressed the button do not like candies.

	Like candies		Does not like candies	
	Red	Green	Red	Green
Pressed button	3	135	22	1
Did not press button	177	60	38	64

    Is liking candy a confounder in this study?

    (A) Yes.

    (B) No.

    (C) There is insufficient information given to determine whether liking candy is a confounder in this study.

    Answer is (B). We have to check whether liking candy is associated with both the colour of the group, and pressing the button.

$$\text{rate}(\text{Like candies} \mid \text{Red}) = \frac{3+177}{3+177+22+38} = 0.75.$$

$$\text{rate}(\text{Like candies} \mid \text{Green}) = \frac{135+60}{135+60+1+64} = 0.75.$$

    Since the two conditional rates are the same, liking candy is not associated to the colour of the group, and is hence not a confounder in this study. Another way to get to this conclusion is to show that

$$\text{rate}(\text{Red} \mid \text{Like candies}) = \text{rate}(\text{Red} \mid \text{Does not like candies}) = 0.48.$$

25. In a certain year, it is known that the prevalence of diabetes among Singapore residents is 10% and the prevalence of diabetes among old (age 60 and above) Singapore residents is 30%. It was suggested that sex is a possible confounder in the observed association between age and diabetes among Singapore residents. After further analysis, the researchers concluded that sex is not a confounder, and there is an association between sex and age. Which of the following statements is/are true? Select all that apply.

    (A) rate(Diabetes | Male) = rate(Diabetes | Female).

    (B) rate(Male | Diabetes) = rate(Female | Diabetes).

    (C) rate(Diabetes | Female) = 10%.

    (A) and (C) are correct. Since sex is not a confounder in the study of association between diabetes and age and sex is associated with age, thus there is **no** association between sex and diabetes. Thus it is correct to say that rate(Diabetes | Male) is equal to rate(Diabetes | Female). Consequently,

    $$\text{rate(Diabetes)} = \text{rate(Diabetes} \mid \text{Male)} = \text{rate(Diabetes} \mid \text{Female)} = 10\%.$$

    On the other hand, we do not know how many diabetic residents are male or female, so we cannot determine if rate(Male | Diabetes) is equal to rate(Female | Diabetes).

26. A researcher would like to find out if there is any relationship between age (young and old) and ramen consumption (high and low) among Singaporeans. From the data he obtained, he suspects that sex is a confounder. Which of the following should hold in order to show that his suspicion is correct? Select all that apply.

    (A) The percentage of old people among males is different from the percentage of old people among females.

    (B) The percentage of males among the high ramen consumers is different from the percentage of females among the high ramen consumers.

    (C) The percentage of high ramen consumers among males is different from the percentage of high ramen consumers among females.

    (A) and (C) are correct. To show that sex is a confounder, we need to show that sex is associated with both age and ramen consumption. If the percentage of old people among males is different from the percentage of old people among females, then sex is associated with age. If the percentage of high ramen consumers among males is different from the percentage of high ramen consumers among females, then sex is associated with ramen consumption. On the other hand, the comparison of percentage of males among high ramen consumers and the percentage of females among high ramen consumers cannot show any association between the variables of ramen consumption and sex.

27. Su is investigating the association between blood pressure and "workaholism" in a certain population. Someone who works more than 75 hours per week is considered a workaholic.

    The income level and blood pressure (high or normal) for each subject and whether or not they are classified as "workaholic" are recorded and summarised in the table below. Here "HBP" denotes "high blood pressure" while "NBP" denotes "normal blood pressure".

	Income Group					
	Low		Middle		High	
	HBP	NBP	HBP	NBP	HBP	NBP
Workaholic	25	75	23	87	26	134
Non-workaholic	25	80	18	72	9	51

    Which of the following statements is true?

(A) We have an instance of Simpson's Paradox for this data set, when considering the association between being a "workaholic" and having "high blood pressure", first for individual income levels ("Low", "Middle", "High") and then overall.

(B) We do not have an instance of Simpson's Paradox for this data set, when considering the association between being a "workaholic" and having "high blood pressure", first for individual income levels ("Low", "Middle", "High") and then overall.

(C) We are not able to determine if we have an instance of Simpson's Paradox for this data set (or not), when considering the association between being a "workaholic" and having "high blood pressure", first for individual income levels ("Low", "Middle", "High") and then overall. There is insufficient information given.

Answer is (A). For "Low", "Middle" and "High" income levels, we have

$$\text{rate(HBP} \mid \text{Workaholic)} > \text{rate(HBP} \mid \text{Non-workaholic)}.$$

For example, for the "Low" income level group,

$$\frac{25}{25+75} = \text{rate(HBP} \mid \text{Workaholic)} > \text{rate(HBP} \mid \text{Non-workaholic)} = \frac{25}{25+80}.$$

But overall, we have

$$\text{rate(HBP} \mid \text{Workaholic)} = \tfrac{(25+23+26)}{(100+110+160)} = 0.2,$$

which is smaller than

$$\text{rate(HBP} \mid \text{Non-workaholic)} = \tfrac{(25+18+9)}{(105+90+60)} = 0.204.$$

28. Some scientists have found that drinking coffee is associated to students' ability to sleep (enough vs not enough sleep). Sex was also found to be a confounder. This means that:

(I) Percentage of coffee drinkers among males is different from the percentage of coffee drinkers among females.

(II) Percentage of males among students who have enough sleep is different from the percentage of males among students who do not have enough sleep.

Which of the statements above is/are correct?

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

Answer is (C). To be a confounder, sex must be associated with both drinking coffeee and sleep. Statement (I) expresses an association between sex and drinking coffee. Statement (II) expresses an association between sex and sleep.

29. A team of researchers are interested in seeing if there is an association between the amount of sleep and memory retention. They have also collected information on each subject's gender. Which of the following statements is correct?

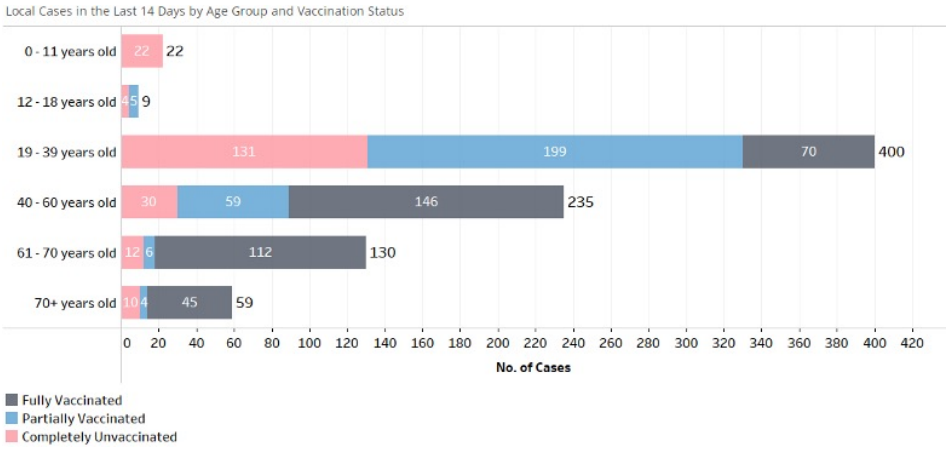
(A) If Simpson's Paradox is not observed when combining the 2 subgroups of gender, then gender is not a confounder when exploring the association between the amount of sleep and memory retention.

(B) Suppose that when the 2 subgroups of gender are combined, Simpson's Paradox is observed when checking for association between the amount of sleep and memory retention. Then gender must be a confounder.

(C) If gender is a confounder when determining the association between the amount of sleep and memory retention, Simpson's Paradox will be observed when combining the 2 subgroups of gender.

Answer is (B). If Simpson's Paradox is observed when data are sliced according to a variable, then the variable must be confounder. The absence of Simpson's Paradox, when data are sliced according to a variable does not mean that the variable is not a confounder.

30.



The above graph, depicting cases of COVID infection (titled "Local cases in the last 14 days by Age group and Vaccination status"), was published in a press release by Singapore's Ministry of Health on 21 July 2021. Let us designate the age range of those 61 years or older as "seniors" , and all others as "non-seniors". Let us also consider the status of either full or partial vaccination as "vaccinated", and the status of being completely unvaccinated as "unvaccinated".

What can be concluded based on the information given? Select all statements that apply.

(A) The rate of infection among seniors is lower than the rate of infection among non-seniors.

(B) For cases depicted by the graphic, being vaccinated is positively associated with seniors.

(C) Age is a confounder in the association between infection and vaccination status.

(D) There were about twelve cases on average daily (correct to the nearest whole number) for those senior and vaccinated.

(B) and (D) are correct. Representing the information from the graph into the following 2-by-2 contingency table (with the stated conditions to form our categories) will make the calculations for the necessary rates much easier.

	Vaccinated	Unvaccinated	Row total
Seniors	167	22	189
Non-seniors	479	187	666
Column total	646	209	855

We are not able to compare the rate of infection among seniors with the rate of infection among non-seniors because we do not have figures for non-infected persons. Thus, we cannot compare rate(Infected | Seniors) vs. rate(Infected | Non-seniors).

However, we can compare rate(Vaccinated | Seniors) vs. rate(Vaccinated | Non-seniors). Using the table:

rate(Vaccinated | Seniors) = 167/189 = 88.36%, which is larger than rate(Vaccinated | Non-seniors) = 479/666 = 71.92%. So being vaccinated and being senior are indeed positively associated with each other.

With reference to the title of the graph, we take the number of cases for those senior and vaccinated (which is 167) and divide it by 14 days. Since 167/14 = 11.9, it is true that there were about 12 cases on average daily (correct to the nearest whole number) for those senior and vaccinated.

To establish that age is a confounder, we need to find that age is associated with both infection and vaccination status. However, as mentioned above, we have insufficient information to show that age and infection are associated with each other. So we cannot determine if age is a confounder in the association between infection and vaccination status.