1 A researcher reported that men have 36 times the odds of drinking wine as compared to women. Which of the following statements must also be true?

A. Men have a higher risk of drinking wine as compared to women
B. Men have a lower risk of drinking wine as compared to women
C. Men have an equal risk of drinking wine as compared to women
D. The relationship between risk of drinking wine among men and risk of drinking wine among women cannot be determined from the information given

2 A researcher aims to examine the rate of Generalised Anxiety Disorder among all undergraduate students in Singapore In order to do so, he obtains a list of all undergraduate students in the National University of Singapore, and randomly selects 1000 students for his sample.
Given that the National University of Singapore is one of various universities in Singapore, which of the following statements is most likely to be true?

A. The researcher has committed the atomistic fallacy
B. The researcher has committed the ecological fallacy
C. The sample is representative of the target population
D. The sample is not representative of the target population

3 In order to determine public attitudes toward various social issues, a newspaper publication firm asks its readers to visit its website to participate in a poll. The poll addresses a different social issue each week. During the previous week, the following poll was found on the website:

Foreign companies should be allowed to participate in domestic political issues.
1 2 3 4 5
Strongly Disagree Disagree Neutral Agree Strongly Agree

The following week, the newspaper publication firm announced that 6 in 10 respondents either disagreed or strongly disagreed with the statement. The population to which the results of this poll can be generalised to is:

A. All readers of the newspaper
B. All readers of the newspaper who have visited the website
C. All readers of the newspaper who have participated in this particular poll
D. All readers of the newspaper who have participated in at least one of the polls

4 In a study in which house addresses were used as the sampling frame, which of the following situations would prevent the researcher from generalising the results of his or her study to a population of citizens and permanent residents?

I. The proportion of addresses with no one living inside of them is 10%
II. The proportion of addresses with no one willing to participate is 80%
III. The sample size is 500

A. I only
B. II only
C. III only
D. II and III only

5 In order to obtain a representative sample of all students in a particular university, a researcher positioned 10 research assistants at the entrance of all of its 10 faculties, and asked them to collect a sample consisting of 1000 opinions by interviewing the first student who entered at the start of each 5 minute mark. Each interview takes about 3 minutes.

After having collected the data, the researcher found that the demographics of the sample matched that of the target population on the following characteristics: age, gender, ethnicity, and major. All of these characteristics were argued to be important in influencing the results of the study.

Which one of the following statements is true?
A. The sample is representative because of its demographic characteristics
B. The sample is representative because of the sampling scheme used to obtain it
C. The sample is not representative because of its demographic characteristics
D. The sample is not representative because of the sampling scheme used to obtain it

6 In 2016, the Singapore population comprised 74% of Chinese ethnicity, 13% of Malay ethnicity, 9% of Indian ethnicity, and 3% of other ethnicities. A researcher stratified the population into these four ethnic groups, and obtained a simple random sample of 200 respondents within each group.

Which of the following statements is true?
A. The sample is representative of the Singapore population because the chance a Chinese individual is selected is the same as that a Malay individual is selected
B. The sample is representative of the Singapore population because the chance a Chinese individual is selected is greater than that a Malay individual is selected
C. The sample is not representative of the Singapore population because the chance a Chinese individual is selected is smaller than that a Malay individual is selected
D. None of the above

7 Among 1000 university students who consent to participate in an experiment, 700 are undergraduates and 300 are postgraduate students. The students are randomly assigned into control group of size 700 and treatment group of size 300.
The number of undergraduates in the treatment group is likely to be ______ the number of postgraduate students in the control group.

A. Less than
B. More than
C. Equal to
D. Cannot be determined

8 A multiple choice exam has 60 questions. Each question has 4 possible answers and only 1 answer out of the 4 possible answers is correct. To receive an A grade, one must answer 95% and above of the questions correctly. We know that 54 questions were answered correctly. What is the probability of receiving an A grade (rounding off to 3 decimal places), if one were to guess the remaining questions?

A. 0.169
B. 0.466
C. 0.500
D. 0.743

9 The probability that a particular coin gets heads is p. Given that $0.4 < p < 0.5$, and that a, b, and c represent the probabilities of the following events:

a: Getting one head and one tail from two coin throws,
b: Getting two heads and one tail from three coin throws, and
c: Getting one head and two tails from three coin throws,

Which one of the following statements must be true?
A. $a < b < c$
B. $b < c < a$
C. $b < a < c$
D. The relationship between a, b and c cannot be determined from the information given

10 The Distress Thermometer (DT) refers to a single-item self-report measure of emotional distress, and is used in outpatient cancer clinics to screen for emotional distress among cancer patients. In a sample of 105 cancer patients, researchers observed that 33 suffered from emotional distress.

After having received a cancer diagnosis, David was screened positive for emotional distress based on the DT. Given that the DT has a sensitivity of 0.88 and a specificity of 0.81, what is the probability that David was actually feeling distressed?
A. 0.88
B. 0.81
C. 0.67
D. 0.94

11 Eric, Freddie and Gavin are three students in a class of 55. A network is created with 55 vertices, and each vertex represents a student in this class. In this network, two vertices are adjacent if the corresponding students have each other's phone numbers, and are not adjacent otherwise. Some of the centrality measures are shown below.

| | Degree Centrality Measure | Closeness Centrality Measure |
|---|---|---|
| Eric | 0.315 | 1.23 |
| Freddie | 0.537 | 1.71 |
| Gavin | 0.611 | 2.34 |

Let N be the number of vertices which are adjacent to both Freddie and Gavin. What is the minimum value that N can take?
A. 2
B. 5
C. 7
D. 9

12 The following adjacency matrix describes the communication network between five soldiers: Alpha (A), Bravo (B), Charlie (C), Delta (D) and Echo (E). In this network, the vertices of two soldiers are adjacent if the corresponding soldiers have direct communication with each other, and are not adjacent otherwise.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | - | 0 | - | - | 1 |
| B | - | - | 1 | - | 0 |
| C | 0 | - | - | 0 | - |
| D | 1 | 0 | - | - | 0 |
| E | - | 0 | 1 | - | - |

Commander Foxtrot decides to choose one of these five soldiers to be her main point of contact, and only considers the closeness centrality measure of this network. Which of the following soldiers is the most appropriate choice?
A. Alpha
B. Bravo
C. Delta
D. Echo

13 A class has 45 students. A network is drawn with 45 vertices, each representing a student in the class. In this network, the vertices of two students are adjacent if the corresponding students are friends, and are not adjacent otherwise. It is known that Alex, Bart and Charles are friends with each other. It is also known that Debbie and Eric are not friends with each other, but are both friends with Alex.
Which one of the following statements must be true about the betweeness centrality measure of Alex?
A. Bcen (Alex) = 0
B. 0 < Bcen (Alex) < 1
C. Bcen (Alex) = 1
D. None of the above

14 An animal shelter has 45 puppies. A network is drawn with 45 vertices, each representing a puppy in the shelter. In this network, the vertices of two puppies are adjacent if the corresponding puppies are friends, and are not adjacent otherwise. It is known that Max has a unique social circle, such that for any two puppies in the shelter excluding Max (e.g., Bella and Charlie), the following equation must be true: d(Bella,Charlie) = d(Bella,Max) + d(Max,Charlie)
Which one of the following statements must be true about the betweeness centrality measure of Max?
A. Bcen (Max) = 0
B. 0 < Bcen (Max) < 1
C. Bcen (Max) = 1
D. None of the above

15 Moving from the current year to the following year, the size of the Age 0 cohort of the following year is calculated based on the data of the current year. Assume that the Age 0 cohort of the following year arises from (a) the live births in the current year, and that
(b) there is no migration at Age 0.
In addition, the following two pieces of information are provided:
1. The total number of births in the current year is 6000, and
2. The sex ratio at birth is 1050.
What is the Age 0 female population of the following year?
A. 2927
B. 3073
C. 3150
D. 6300

16. Moving from the current year to the following year, the size of the Age X cohort* of the following year is calculated based on the data of the current year. Assume that the Age X cohort of the following year arises from (a) the Age X-1 cohort in the current year, minus (b) the Age X-1 deaths in the current year, and plus (c) the Age X migrants. * X is between 1 and 84 inclusive

In addition, the following three pieces of information are provided:
1. The Age X-1 cohort of the current year comprises 8000 individuals,
2. The Age X-1 death rate of the current year is 25, and
3. The Age X migration rate is 45.

What is the Age X population of the following year?
A. 7615
B. 7845
C. 8151
D. 8160

17. Moving from the current year to the following year, the size of the Age >85 cohort of the following year is calculated based on the data of the current year. Assume that the Age >85 cohort of the following year arises from (a) the Age >84 cohort in the current year, minus (b) the Age >84 deaths, and plus (c) the Age >85 migrants.

In addition, the following three pieces of information are provided:
1. The Age 84 and Age >85 cohorts of the current year comprise 3000 and 10000 individuals respectively,
2. The Age 84 and Age >85 death rates of the current year are 55 and 105 respectively, and
3. The Age 84 and Age >85 migration rates of the current year are -2 and -5 respectively.

What is the Age >85 population of the following year?
A. 11577
B. 11726
C. 11735
D. 11844

18. In a peculiar town where everyone is above 19 years old and retires at 70 years old, the old age support ratio among males is two, and the old age support ratio among females is one. What are the odds x that an individual in this town is above the age of 69 years?
A. 0.50 ≤ x ≤ 0.67
B. 0.33 ≤ x ≤ 0.50
C. 0.50 ≤ x ≤ 1.00
D. 0.67 ≤ x ≤ 0.75

1. A
2. D
3. C
4. B
5. D
6. D
7. C
8. A
9. B
10. C
11. C
12. D
13. B
14. C
15. A
16. C
17. B
18. C

1. Among 1000 university students who consent to participate in an experiment, 700 are undergraduates and 300 are postgraduate students. The students are randomly assigned into control group of size 700 and treatment group of size 300. The number of undergraduates in the treatment group is likely to be the number of postgraduate students in the control group.
(A) less than
(B) equal to
(C) more than

2. A researcher reported that men have 36 times the odds of drinking wine as compared to women. Which of the following statements must also be true?
(A) Men have a higher risk of drinking wine as compared to women
(B) Men have a lower risk of drinking wine as compared to women
(C) Men have an equal risk of drinking wine as compared to women
(D) The relationship between risk of drinking wine among men and risk of drinking wine among women cannot be determined from the information

3. A researcher aims to examine the rate of Generalised Anxiety Disorder among all undergraduate students in Singapore In order to do so, he obtains a list of all undergraduate students in the National University of Singapore, and randomly selects 1000 students for his sample. Given that the National University of Singapore is one of various universities in Singapore, which of the following statements is most likely to be true?
(A) The researcher has committed the atomistic fallacy
(B) The researcher has committed the ecological fallacy
(C) The sample is representative of the target population
(D) The sample is not representative of the target population

4. In order to determine public attitudes toward various social issues, a newspaper publication firm asks its readers to visit its website to participate in a poll. The poll addresses a different social issue each week. During the previous week, the following poll was found on the website: Foreign companies should be allowed to participate in domestic political issues.
1 : Strongly disagree
2 : Disagree
3 : Neutral
4 : Agree
5 : Strongly agree

The following week, the newspaper publication firm announced that 6 in 10 respondents either disagreed or strongly disagreed with the statement. The population to which the results of this poll can be generalised to is:
A. All readers of the newspaper
B. All readers of the newspaper who have visited the website
C. All readers of the newspaper who have participated in this particular poll
D. All readers of the newspaper who have participated in at least one of the polls

5. In order to obtain a representative sample of all students in a particular university, a researcher positioned 10 research assistants at the entrance of each of the 10 faculties (the university has a total of 10 faculties), and asked them to collect a sample consisting of 1000 opinions by interviewing the first student who entered at the start of each 5 minute mark (interview takes about 3 minutes).

After having collected the data, the researcher found that the demographics of the sample matched that of the target population on the following characteristics: age, gender, ethnicity, and major. All of these characteristics were argued to be important in influencing the results of the study. Which of the following statement is true?
(A) The sample is representative because of its demographic characteristics
(B) The sample is representative because of the sampling scheme used to obtain it
(C) The sample is not representative because of its demographic characteristics
(D) The sample is not representative because of the sampling scheme used to obtain it

6. In a study in which Singapore house addresses were used as the sampling frame, which of the following situations would prevent the researcher from generalising the results of his/her study to the population of Singapore?
(i) The proportion of addresses with no one living inside of them is 10%.
(ii) The proportion of addresses with no one willing to participate is 80%.
(iii) The sample size is 500.
(A) (i) only
(B) (ii) only
(C) (iii) only
(D) (ii) and (iii) only

7. In 2016, the Singapore population comprised 74% of Chinese ethnicity, 13% of Malay ethnicity, 9% of Indian ethnicity, and 3% of other ethnicities. A researcher stratified the population into these four ethnic groups, and obtained a simple random sample of 200 respondents within each group. Which of the following statements is true?

(A) The sample is representative of the Singapore population because the chance a Chinese individual is selected is the same as that a Malay individual is selected

(B) The sample is representative of the Singapore population because the chance a Chinese individual is selected is greater than that a Malay individual is selected

(C) The sample is not representative of the Singapore population because the chance a Chinese individual is selected is smaller than that a Malay individual is selected

[✓ (D)] None of the above

8. Which of the following is/are true about p-value?

(i) P-value is a conditional probability computed based on the assumption that null hypothesis is true.

(ii) P-value gives the probability that null hypothesis is true.

(iii) P-value is dependent on the sample size of the hypothesis test.

(iv) A small p-value provides evidence that the null hypothesis is not true.

(v) A large p-value provides evidence that the null hypothesis is true.

(A) (i), (ii), (iii), (iv) and (v) only

(B) (i), (ii), (iv) and (v) only

(C) (i), (iii), (iv) and (v) only

[✓ (D)] (i), (iii) and (iv) only

9. A multiple choice exam has 60 questions. Each question has 4 possible answers and only 1 answer out of the 4 possible answers is correct. To receive an A grade, one must answer 95% and above of the questions correctly. We know that 54 questions were answered correctly. What is the probability of receiving an A grade (rounding off to 3 decimal places), if one were to guess the remaining questions?

[✓ (A)] 0.169

(B) 0.466

(C) 0.500

(D) 0.743

10. A biased coin has probability p of getting heads and suppose $0.4 < p < 0.5$.

Let a, b and c be the probabilities of the following events:

a: Getting one head and one tail from two coin throws.

b: Getting two heads and one tail from three coin throws.

c: Getting two tails and one head from three coin throws.

Which of the following is true?

(A) $a < b < c$

[✓ (B)] $b < c < a$

(C) $b < a < c$

(D) The relationship between a, b and c cannot be determined from the information given

11. In a certain day care class, 30% of the children have brown eyes, 20% of them have blue eyes and the other 50% have eyes that are in other colors. One day some of them play a game together. In the game, 45% of the children have brown eyes, 20% have blue eyes and 35% have other eye colors. Now, if a child is selected randomly from the class, and we know that he/she was not in the game, what is the probability that the child has blue eyes (rounding off to 2 decimal places)?

(A) 0.00

(B) 0.04

(C) 0.11

[✓ (D)] 0.20

12. Assuming a telemarketer has a 20% chance of selling item A to each caller, a 40 % chance of selling item B to each caller, and the event of selling item A is independent to the event of selling item B. Each call in which the telemarketer sells both item A and item B takes 150 seconds, each call that sells only one item (item A or item B but not both) takes 60 seconds and each call that doesn’t sell any items takes 30 seconds. If the telemarketer makes 45 calls, what is the average amount of time it takes?

(A) 53 seconds

(B) 2268 seconds

[✓ (C)] 2376 seconds

(D) 2808 seconds

13. There is an epidemic. A person has probability 0.01 of having the disease. The authorities decide to test the population, but the test is not completely reliable. The sensitivity of the test is 0.98 and the specificity of the test is 0.95. Patrick was tested positive for the disease, what is the probability that Patrick has the disease (rounding off to 3 decimal places)?
(A) 0.010
(B) 0.020
(C) 0.165
(D) 0.198

14. Referring to Question 7, Patrick wants a second opinion: an independent repetition of the test (regardless of Patrick's disease status, outcomes of tests are independent). He went for a second test and was tested positive again. What is the probability that Patrick has the disease?
(A) 0.027
(B) 0.165
(C) 0.795
(D) 0.960

15. Eric, Freddie and Gavin are three students in a class of 55. A network is created with 55 vertices, and each vertex represents a student in this class. In this network, two vertices are adjacent if the corresponding students have each others phone numbers. Some of the centrality measures are shown below.

| | Degree centrality measure | Closeness centrality measure |
|---|---|---|
| Eric | 0.315 | 1.23 |
| Freddie | 0.537 | 1.71 |
| Gavin | 0.611 | 2.34 |

Let N be the number of vertices which are adjacent to both Freddie and Gavin. What is the minimum value that N can take?
(A) 2
(B) 5
(C) 7
(D) 9

16. Suppose you are going to open a new airline which operates a network of flights between 6 cities. In the network, a vertex represents a city, and two vertices are adjacent if there is a direct flight between the two cities.
You must ensure that any city can be reached from any other city either directly or via one connecting city. Also, due to the limited start-up fund, the size of the graph should be as small as possible. There are a few graphs satisfying the above requirements. Among all these graphs, what is the smallest (possible) closeness centrality measure of the cities?
(A) 1
(B) 1.2
(C) 1.4
(D) 1.6

17. A particular network has 4 vertices. There is one vertex u where there is exactly two vertices at distance 1 from u, and exactly one vertex at distance 2 from u. Moreover, all vertices have degree greater than 1.
Which of the following statement is correct?
(A) The degree of u is 3
(B) It is not possible to determine which vertex in the network has the smallest closeness centrality measure based on the information given in the question
(C) It is not possible to determine which vertex in the network has the largest closeness centrality measure based on the information given in the question
(D) None of the above

18. A class has 45 students. A network is drawn with 45 vertices, each representing a student in the class. Two vertices are joined directly by an edge if the two corresponding students are friends. It is known that student A, student B and student C are friends with each other. Suppose there are two other students in class, student D and student E, who are not friends with each other but both are friends of student A.
Which of the following is true about the betweenness centrality measure of student A?
(A) Bcen (student A) = 0
(B) 0 < Bcen (student A) < 1
(C) Bcen (student A) = 1
(D) None of the above

~~19~~. Another class has 43 students. The network of this class is drawn according to the criteria in Question 18. It is known that a student in this class, student H, has a unique social circle. For any two students in this class (excluding student H), say student S and student T, we have d(student S, student T) = d(student S, student H) + d(student H, student T).

Which of the following is true about the betweenness centrality measure of student H?
(A) Bcen (student H) = 0
(B) 0 < Bcen (student H) < 1
(C) Bcen (student H) = 1
(D) None of the above

~~20~~. With reference to the discussion of the Bacon number of an actor, suppose the Bacon number of an actor A is 2. It is known that actor A, actor B and actor C were in the same movie "Frank". Moreover, actor A and actor C acted in only one movie.

Which of the following statements must be correct?
(i) Bacon number of actor A is less than bacon number of actor B.
(ii) Degree of actor B is less than degree of actor C.
(iii) Closeness centrality measure of actor A is the same as the closeness centrality measure of actor B.
(iv) Betweenness centrality measure of actor A is the same as the betweenness centrality measure of actor C.

(A) (iv) only
(B) (i) and (ii) only
(C) (iii) and (iv) only
(D) None of the above

~~21~~. With reference to the discussion of the Bacon number of an actor, suppose the Bacon number of an actor C was initially computed to be 3. However, it was later discovered that there was one error in the database. The error was namely:

(i) There is a movie "Happiness", not involving Kevin Bacon, that was omitted from the database.

Suppose we know that everyone who acted in the movie "Happiness" has an initial Bacon number of at least 2. Which of the following statement must be correct after the error was corrected?
(A) The Bacon number of actor C is now 1
(B) The Bacon number of actor C is now 2
(C) The Bacon number of actor C is still 3
(D) It depends, the Bacon number of actor C now could be 2 or 3

~~22~~. In 2013, the sex ratio at birth of a certain country is 900. In 2014, the death rate of 0-year-old males is 10 per 1000 and the death rate of 0-year-old females is 7 per 1000. The 2015 sex ratio at 1 year old is (assuming there is no age 1 migration)
(A) less than 890
(B) between 890 to 895
(C) between 895 to 900
(D) more than 900
(E) We cannot determine based on the given information

~~23~~. In a peculiar town where everyone is above 19 years old and retires at 70 years old, the old age support ratio (OASR(70)) among males is two, and the old age support ratio (OASR(70)) among females is one. What are the odds x that an individual in this town is above the age of 69?

(A) 0.50 ≤ x ≤ 0.67
(B) 0.33 ≤ x ≤ 0.50
(C) 0.50 ≤ x ≤ 1.00
(D) 0.67 ≤ x ≤ 0.75

~~24~~. In 2015, OASR(75) = ½ and OASR(76) = 2/3.

What is the rate (Age 75 population | Age ≥ 75 population) in 2015?
(A) 1/3
(B) 1/5
(C) 1/10
(D) 2/3

25. The age specific fertility rates (ASFR) for the year 2013 are given as follow:

| Age group | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 |
|---|---|---|---|---|---|---|---|
| ASFR | 8.00 | 30.20 | 70.80 | 86.20 | 30.50 | 6.30 | 0.70 |

Suppose it is known that between 2013 and 2014, ASFR increases by 4% per year for the age group 20-24 and age group 30-34 while it increases by 2% per year for the age group 15-19 and age group 35-39. ASFR remains unchanged for all other age groups. Rounding off to 2 decimal places, the total fertility rate (TFR) for the year 2014 is

A. 1.15
B. 1.17
C. 1.18
D. 1.19

26. Assume that the Age X cohort of the following year arises from those who are Age (X-1) in the current year, minus deaths of Age (X-1), plus net migration of Age X. Given that:

- the Age (X-1) cohort of the current year is 740,
- the Age (X-1) deaths of the current year is 42,
- the Age X migration rate is 24 per 1000, the Age X population (rounded off to whole number) for the following year is

(A) 715
(B) 716
(C) 722
(D) 727

27. Moving from the current year to the following year, the size of the Age ≥ 85 group in the following year is computed based on the data of the current year. In our model, we assume that the Age ≥ 85 group arises from the Age ≥ 84 group in the current year, minus deaths (of Age ≥ 84), plus migration (of Age ≥ 85). Given that:

- The Age 84 and Age ≥ 85 cohorts of the current year comprise 3000 and 10000 individuals respectively,
- The Age 84 and Age ≥ 85 death rates of the current year are 55 and 105 respectively, and
- The Age 84 and Age ≥ 85 migration rates of the current year are -2 and -5 respectively.

What is the Age ≥ 85 population of the following year?

(A) 11577
(B) 11726
(C) 11735
(D) 11844

28. Suppose Age X population in 2015 is smaller than the Age (X-1) population in 2014. We also know that Age X death rate in 2015 is smaller than Age (X-1) death rate in 2014 and Age (X+1) migration rate in 2015 is larger (in value) than Age X migration rate in 2014. Which of the following statements about Age (X+1) population in 2016 must be correct?

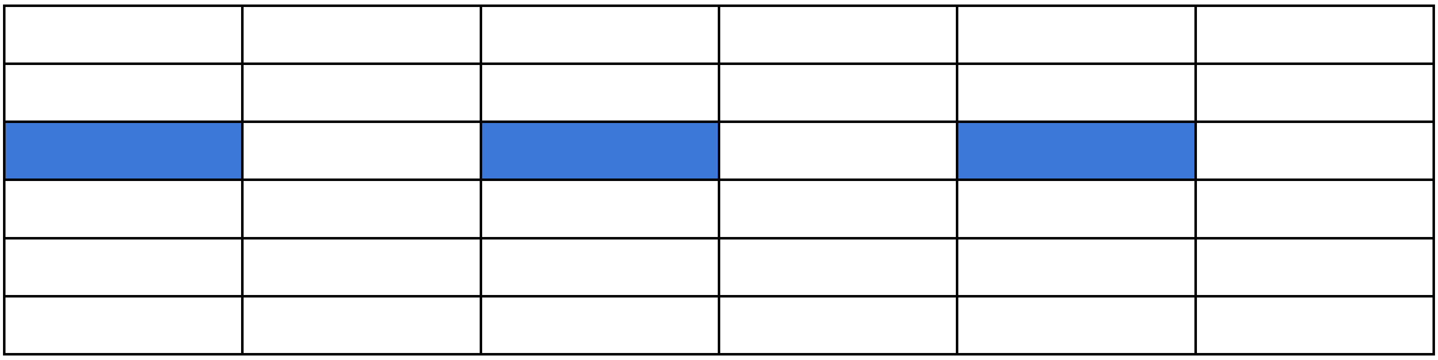
(i) It is larger than Age X population in 2015.
(ii) It is smaller than Age X population in 2015.
(iii) It is larger than Age (X-1) population in 2014.
(iv) It is smaller than Age (X-1) population in 2014.

(A) (i) only
(B) (i) and (iii) only
(C) (ii) and (iv) only
(D) None of the above

1. B
2. A
3. D
4. C
5. D
6. B
7. D
8. D
9. A
10. B
11. D
12. C
13. C
14. C
15. C
16. A
17. B
18. B
19. C
20. A
21. C
22. C
23. C
24. C
25. D
26. A
27. B
28. D

1. Which of the following statements about outliers are true?
(I) We should usually, but not always, remove outliers from the data collected.
(II) Removal of outliers will understate the strength of the correlation.
A. Only (I) is true.
B. Only (II) is true.
C. Both (I) and (II) are true.
[✓ D. Neither (I) nor (II) are true.]

2. Which of the following sampling plans are best described by the diagram below?



A. Simple Random Sampling
[✓ B. Systematic Sampling]
C. Volunteer Sampling
D. Non-probability Sampling

3. A researcher, Dr Poso, is conducting a study to investigate if watching violent cartoons improves cognitive function in children. He decides to survey parents of young children. Which of the following situations could cause bias in the results of a survey where house addresses are used as the frame?
(I) A selected address in the sample is unoccupied. It has no residents living there.
(II) The person who opened the door refused to participate.
(III) There was no answer, so the surveyor gave up and visited the next nearest neighbour on the right.

A. (I), (II) and (III) all cause bias.
B. Only (I) and (II) cause bias.
C. Only (I) and (III) cause bias.
[✓ D. Only (II) and (III) cause bias.]

4. The National Automobile Association's (NAA) Used Car Buyer's Guide is "a compilation of consumer-oriented information designed to reduce the frustration and uncertainty often associated with buying a car." A questionnaire is distributed to all NAA members, motorist through consumer magazines, random mailings, newspaper ads, and public press release. A total of 15,446 responses were received and analyzed this year. A reporter is doing a story on the quality of cars, and will use NAA's survey.

Based on the above information, which of the following statements are true?
A. The frame used in this survey is a good frame, as it covers a wide spectrum of motorists.
B. The size of the sample is large. It should have a good representation of car owners in the population.
[✓ C. A non-random sampling plan was used in the survey.]
D. None of the above.

5. How does Vitamin D levels affect the risk of Cognitive decline among Chinese Elderly people? Data consisting of people from eight longevity areas in China were used. Suppose the researchers wish to extend this study to the population living in rural remote areas, and that it requires at least two days to travel from one remote area to the next. If the researchers are on a limited budget, which of the following sampling plans would you recommend?
(I) Multi-stage Sampling
(II) Cluster Sampling
(III) Simple Random Sampling
[✓ A. Only (I) and (II)]
B. Only (I) and (III)
C. Only (II) and (III)
D. (I), (II), and (III)

6. To investigate the overall satisfaction of bus commuters, a surveyor stations herself at a bus interchange. She interviews the third person who walks past her. When she is done with the interview, proceeds to interview the next third person who walks past her. She continues this until she has 100 responses. What sampling plan was used?
A. Multi-stage Sampling
B. Quota Sampling
C. Judgement Sampling
[✓ D]. None of the above

7. An airline company requires crew members to submit themselves to urinalysis for drug detection before each flight. Among drug users, 95% will test positive, and among drug-free persons, 95% will test negative. Let us assume that 5% of airline crew take drugs from time to time. Compute the conditional probability that a crew member is a drug user given that he is detected positive.
A. 0.05
B. 0.00028
C. 0.95
[✓ D]. 0.5

8. An insurance company charges $800 per year per customer for a certain health insurance policy with a maximum payout of $20,000 when a customer makes a claim. Each year, 3% of the customers submit a claim. Among them, 50% of these customers get a payout of $5000, while the remaining 50% get the maximum payout of $20,000.
What is the average gain to the insurance company per customer who buys this policy?
A. $125
B. $200
C. $375
[✓ D]. $425

9. A policeman was quoted as saying that the proportion of a certain ethnic minority among those convicted of robbery was higher than the proportion in a general population. Which of the following can be made from the policeman's statement?

(I) A randomly selected member of the ethnic minority is more likely to be convicted of robbery than a random member of the population.
(II) A randomly selected member of the ethnic minority is likely to be convicted of robbery.

[✓ A]. Only (I)
B. Only (II)
C. Both (I) and (II)
D. Neither (I) nor (II)

10. Mr. J recently bought a 20-sided dice. Each side was labelled 1 to 20. The shop owner told him that it was manufactured with such precision that it had every side has the exact same chance of showing up. After rolling the dice 5 times, Mr. J was puzzled to discover that it had landed on a ‘7’ 4 out of 5 times. Perform a hypothesis test. Is the dice biased?

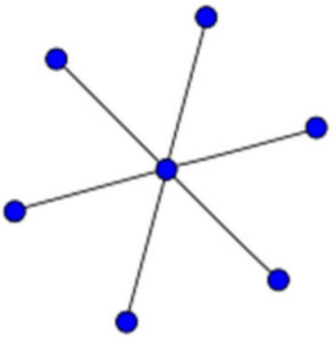
A. The null hypothesis is that the dice is biased, and we conclude that it is indeed biased at a 1% level of significance.
[✓ B]. The null hypothesis is that the dice is not biased, but we conclude that it is biased at a 1% level of significance.
C. The null hypothesis is that the dice is biased, but we conclude that the dice is not biased at a 1% level of significance.
D. The null hypothesis is that the dice is not biased, but no conclusion can be drawn.

11. In a certain country, the proportion of residents with blood type A is 0.41. From the above information, which of the following statements are correct?

(I) Joy is a resident in this country. The probability that she has blood type A is 0.41.
(II) A person is selected at random from the resident population. The probability that he/she has blood type A is 0.41.

A. Only (I)
[✓ B]. Only (II)
C. Both (I) and (II)
D. Neither (I) nor (II)

12. A star graph is a network where one of its vertices, u, is adjacent to all the other vertices. Furthermore, all vertices other than u are only adjacent to u. As an example, a star graph of 7 vertices is shown below:



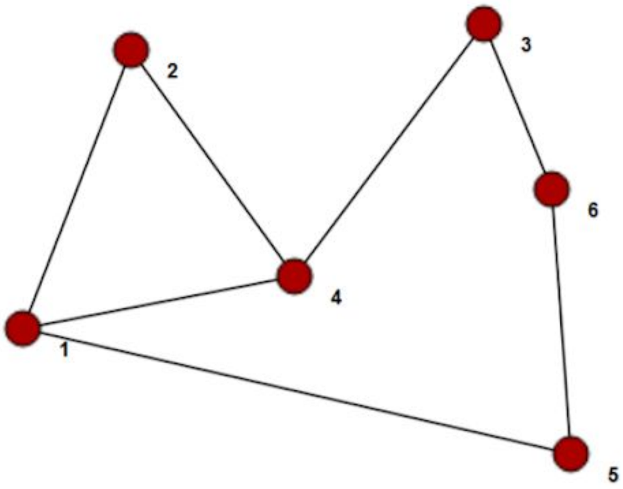
Now consider a star graph of 20 vertices. Which of the following statements are true?
(I) The Bcen of all 20 vertices are either 0 or 1.
(II) The Ccen of all 20 vertices are either 0 or 1.
A. Only (I)
B. Only (II)
C. Both (I) and (II)
D. Neither (I) nor (II)

13. The adjacency matrix of a network of 5 vertices is given below:

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 0 |
| 2 | 1 | 0 | 1 | 1 | 1 |
| 3 | | | | | |
| 4 | 1 | 1 | 0 | 0 | 0 |
| 5 | 0 | 1 | 1 | 0 | 0 |

Which should be the values for the blank row?
A. 1,1,0,1,0
B. 0,1,1,0,1
C. 0,1,0,0,1
D. 1,1,1,0,1

14. The following diagram represents a computer communications network. Each vertex represents a computer device, and two devices are adjacent if they can exchange data with each other using a data link.



Suppose that we wish to disseminate a piece of information to all devices in that network in the shortest time possible. Assuming that the time taken to transfer data between all adjacent devices are identical, which device (or vertex) will you choose as the source to broadcast this information?
Choose the best option.
A. I will choose vertex 1, because of its Betweenness centrality measure
B. I will choose vertex 1, because of its Closeness centrality measure
C. I will choose vertex 3, because of its Betweenness centrality measure
D. I will choose vertex 3, because of its Closeness centrality measure

15. With reference to the movie graph, recall that the Bacon number of an actor is defined to be the distance from Kevin Bacon. Also, 2 actors are adjacent if they have acted in the same movie. At the end of 2014, the Bacon number of an actor A was 3. In 2015, neither Kevin Bacon nor actor A acted in any movie.
Which of the following can happen at the end of 2015?
(I) The Bacon number of actor A is 1.
(II) The Bacon number of actor A is 2.
(III) The Bacon number of actor A is 3.
(IV) The Bacon number of actor A is 4.

A. (III) only.
B. (II) and (III) only.
C. (III) and (IV) only.
D. (I), (II) and (III) only.
E. (I), (II), (III) and (IV).

16. A network has an order of 500,000, and the degree of every vertex in the network is 10.
Suppose x is a vertex in the network. How many vertices are at distance 2 from x?
(I) It is possible that there are 9 vertices at distance 2 from x.
(II) It is possible that there are 90 vertices at distance 2 from x.
(III) It is possible that there are 100 vertices at distance 2 from x.

A. Only (I) and (II) are true
B. Only (I) and (III) are true
C. Only (II) and (III) are true
D. (I), (II), and (III) are true
E. None of the above.

17. Recall that the Old-Age Support Ratio (OASR) for a population is defined as

$$OASR = \frac{\#\ aged\ 20\ to\ 64}{\#\ aged \geq 65}.$$

A demographer Bill calculated the OASR of country X, and it was found to be 6.2.
However, Bill notes that X is a less-developed country, and a significant proportion of the working population are below 20 years of age. He therefore suggests to replace the OASR by a more suitable measure for country X, the New Support Ratio (NSR):

$$NSR = \frac{\#\ aged\ 15\ to\ 64}{\#\ aged \geq 65}.$$

What can be said about the NSR for country X?
A. NSR < 6.2
B. NSR > 6.2
C. NSR = 6.2
D. It cannot be determined if the NSR is less or more than 6.2, as more information is needed.

18. You have been engaged by the government of Pulau Guleam to project the number of 75-yearolds for the year 2018. Pulau Guleam's last census was taken in 2016, and the government has provided you with the census data below. It is known that death and migration rates in Pulau Guleam have been stable in the last 10 years, and are not expected to change in the next several

| Years. | 70 | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 |
|---|---|---|---|---|---|---|---|---|---|---|
| Population | 11879 | 10726 | 9572 | 8944 | 8317 | 7971 | 7343 | 6716 | 6146 | 5577 |
| Death rate | 21.2 | 23.5 | 25.8 | 29.2 | 32.7 | 36.3 | 39.7 | 43.2 | 48.2 | 53.2 |
| Migration rate | 12.2 | 11.4 | 9.8 | 7.7 | 7.5 | -4.7 | 0 | 0 | 0 | 0 |

Using the given 2016 census data, project of the number of 75-year-olds in the year 2018.
A. 8543
B. 8422
C. 8748
D. 8007
E. Cannot be determined, as the birth rates are not given.

19. A demographer would like to know the Age-Specific Fertility Rates (ASFRs) of a certain country. He looks up the national database and finds the following data:

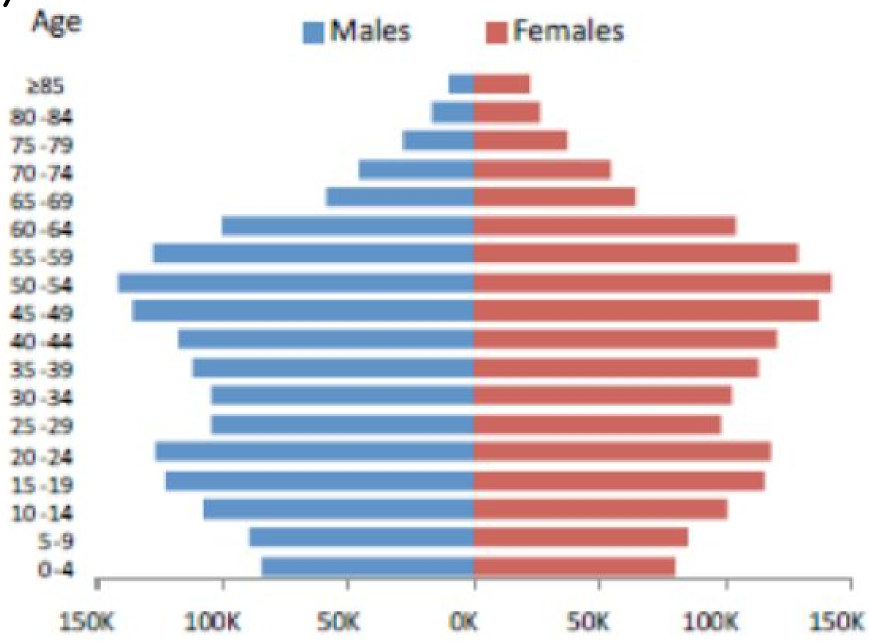
| Age | Fertility Rate |
|---|---|
| 15 | 2.468 |
| 16 | 2.417 |
| 17 | 4.734 |
| 18 | 4.697 |
| 19 | _3# |

Unfortunately, a hacker had gained access to the database, and changed the age-19 ASFR to "_3#". The demographer is in great distress. He searches through the written records and finds that the ASFR for the age group of 15-19 years is 4.258.

From the above information, what can be said about the age-19 ASFR? Choose the best option.
A. It can be calculated exactly, and it is 6.974.
B. It cannot be calculated exactly, but it can be estimated to be 4.258.
C. It cannot be calculated or estimated. We need to know the ASFRs for all childbearing ages
from 15 to 49 years.

20. The Sex Ratio at birth for a certain country was 1070 in the year 2015. On closer inspection of the census data, it was found that 5% of male births in 2015 were actually female births.
What should be the Sex Ratio at birth in 2015?
A. 1017
B. 965
C. We cannot calculate the exact value, but it is less than 1070
D. It should still be 1070

21. The population pyramid for year 2013 of a certain country shown below:



If fertility rates were to remain constant from 2013 to 2023, which of the following are true?
A. The number of working adults will remain constant for the years 2013 to 2023.
B. The number of retirees will increase, for the years 2013 to 2023.
C. The number of births will decrease, for the years 2013 to 2023.
D. None of the above.

1. Ans: D. (I) Outliers should be handled with care. (II) Depends. Outliers can either increase, or decrease the strength of the correlation.
2. B
3. Ans: D. (II) could cause bias due as those who do not respond may have a different opinion from those who do. (III) could cause bias, for reasons similar to (II). (I) does not cause bias as no population unit is excluded. The sampling frame can be larger than the population
4. Ans: C.
   The target population is all used car owners. The sampling frame is not well defined and therefore the sample is not well selected. The eventual sample is derived from NAA members, random mailings, and volunteer sample (motorist through consumer magazines, newspaper ads and public press release). It does not matter how large the size of the responses is. It is a bad sample and there is no way to ascertain whether or not it represents any larger group.
5. Ans: A. Cluster sampling is useful if each cluster is a geographical location (eg. a country). Surveyors / Interviewers only need to travel to these selected locations, which reduces the expensive transport costs (and time) than if they were to travel to each individual location. Multi-stage sampling is similar to cluster sampling, except that a (sub-)sample is chosen from each cluster.
6. Ans: D. It is not A as non-random sampling was used.
   Even though there is a target of 100 responses to meet, this is not B (quota sampling). Quota sampling involves a quota for multiple groups, or categories, based on a certain characteristic (eg. a quota to meet for each of the different races).
   Neither is it C (judgement sampling). If it was so, she would have used her own judgement to decide who is representative of the population.
7. Ans: D. Assume 100,000 tests are administered (you may change the total number of tests). Use the information given to fill in the 2x2 table:

   | | tested positive | tested negative | Row sum |
   |---|---|---|---|
   | drug users | 4,750 | 250 | 5,000 |
   | drug-free | 4,750 | 90,250 | 95,000 |
   | Column sum | 9,500 | 90,500 | 100,000 |

   P(drug user | tested positive) = 4,750/9,500 = ½.
   [Remark] The reason that this probability is much lower than the sensitivity of the test is because of a low base rate the base rate of 0.05.
8. Ans: D. Expected profit = $800 * 97% + -$4200 * 1.5% + $(800-20,000) * 1.5% = $425.
9. Ans: A. Only (I). 'Translate' the question into symbols. The policeman's statement says that P(M|R) > P(M). By the consistency rule, we have P(R|M) > P(R). This is statement (I).
   Statement (II) is false, as the events R, and M are dependent. ( P(R|M) is not equal to P(R) )
   [Also, refer to Q11.]
10. Ans: B. The null hypothesis should be that the dice is fair, or not biased.
    p-value = P(dice lands 4 or more times on a 7, given that the dice is not biased) = P(dice lands 4 times on 7, given that the dice is fair) + P(dice lands 5 times, given that… ) = 5 x (1/20)^4 x (19/20) + (1/20)^5 = 0.00003 < 0.01.
    (The p-value is the probability that 4 or more '7's are observed, if the dice is fair).
11. Ans: B. Explanation: this proportion of 0.41 applies to the population as a whole, and may differ from person to person. Refer to Tut 3 Q2(c) (about the tennis club).
12. Ans: A. The Bcen of the central vertex u is 1, since any pair of points has u in between them. The Bcen of any 'outer vertex' is 0, since any pair of points will not need to pass through that point. Thus (I) is correct.
    The Ccen of the central vertex is 1, but the Ccen of the 'outer vertices' are more than 1. Thus (II) is incorrect. [Actually, Ccen is always 1 or more]
13. Ans: C. Note that the adjacency matrix is symmetric along the diagonal. For example, the value in the 1 st row, 3rd column should be the same as the value in the 3rd row, 1st column. (If vertex 1 and 3 are adjacent, this also means that vertex 3 and 1 are adjacent.) Also, the values along the diagonal (1st row, 1st column; 2nd row, 2nd column etc.) are always 0.
    [Optional, for interest] Note that all networks in this module are simple graphs, so that the adjacency matrix consists only of 0's and 1's. This may not be the case for more complex graphs.
14. Ans: B. In this case, we should consider the closeness centrality measure.
    Vertex 1 (and also vertex 4) has the lowest Ccen.

    | Node | Bcen | Ccen |
    |---|---|---|
    | 1 | 0.25 | 1.4 |
    | 2 | 0 | 1.8 |
    | 3 | 0.15 | 1.6 |
    | 4 | 0.25 | 1.4 |
    | 5 | 0.15 | 1.6 |
    | 6 | 0.1 | 1.8 |

15. Ans: A. Draw a sketch.
    Since both Kevin Bacon and Actor A did not act in a movie together in 2015, (I) cannot be correct. Since Actor A did not act in a movie with any other actor with Bacon number 1 in 2015, (II) cannot be correct. The Bacon number of Actor A cannot increase from 3 to 4 since there were no removal of any vertices or edges in the movie network, so (IV) cannot be correct.
16. Ans: A. (This is a difficult question)
    Explanation: (II) is possible, when x is adjacent to 10 vertices, and each of these 10 vertices are adjacent to x and 9 other vertices. In total, there will be 9 x 10 = 90 vertices at distance 2 from x.
    Draw it out to see this. (I) is also possible. x is adjacent to 10 vertices, and each of these 10 vertices are adjacent to x and the same 9 vertices. (ie. these 10 vertices “share” 9 other neighbours). Then there will be only 9 vertices at distance 2 from x.
    Draw it out to see this. (III) is not possible, as vertices adjacent to x can only be adjacent to 9 other vertices apart from x.
17. Ans: B. The numerator for NSR contains a larger group, compared to the numerator for OASR.
18. Ans: B. Start with age 73 population and project it 2 years forward. 8944*(1-29.2/1000)*(1+7.5/1000) = 8748 will be the projected age 74 population in 2017. 8748*(1-32.7/1000)*(1-4.7/1000) = 8422 will be the projected age 75 population in 2018. Birth rates are not needed, as new births will be age 0 in the next year.
    Note that age 74 migration rates were used when projecting the age 74 population. Similarly when projecting the age 75 population. [refer to summary slides for details]
19. Ans: B is the best option. This estimate is good if the number of women in each individual age, from 15-19, are roughly equal. [4.258 will be the (simple) average of the 5 individual ASFRs if the number of women for each childbearing age from 15-19 years were equal. Otherwise, the overall average (4.258) depends on the proportion of women in each age.]
    In order to calculate the ASFR for age-19, the number of women for each childbearing age from 15 to 19 years needs to be known.
20. Ans: B. The ratio of male births to female births was initially 1070/1000. The proportion of male births out of the total number of births was initially 1070/2070.
    With the 5% change, the proportion of male births would be 95% * 1070/2070 = 1016.5/2070. Note that the denominator 2070 does not change [Why?]
    Therefore the sex ratio at birth will be 1016.5 / (2070 - 1016.5) = 1016.5 / 1053.5, which is 964.9 / 1000.
21. Ans: D. None of the above. A population pyramid only tells us the demographic profile of its population in that year. It does not mention anything about the past or future. (Death, migration rates, and other factors may change.) More than one population pyramid should be used if a trend is to be studied.

Risk and Odds

Suppose in a population of size n, s people have a disease. The risk of the disease is r = s/n, and the odds of the disease is r/(1-r) which is the same as s/(n-s).

1. DES was given to pregnant women to prevent miscarriage. A literature review found 3 randomised controlled experiments and 5 nonrandomised studies with control groups. The rate of miscarriages was about the same in all the 8 treatment groups and in the control groups in the 3 randomised controlled experiments. However, the rate was substantially higher among the control groups in the 5 nonrandomised studies.

A. DES is effective.  
B. DES is not effective.  
C. The result is inconclusive.

2. In a US General Social Survey, 2,726 adults were asked about education level and feelings about the future.  
The responses are as follows:

| | Optimistic | Pessimistic | Row sum |
|---|---|---|---|
| Without high school diploma | 748 | 1336 | 2084 |
| With high school diploma | 138 | 504 | 642 |
| Column sum | 886 | 1840 | 2726 |

Which of the following justifies the statement “Among the survey participants, not having high school diploma was associated with optimism about the future.”?  
(I) 1336/2084 > 748/2084  
(II) 748/2084 > 138/642  
(III) 748/886 > 1336/1840

A. Only (I).  
B. Only (II). C. Only (III).  
D. Only (II) and (III).  
E. All of them.

3. Since the 1950’s, many observational studies found consistent association between high- density lipoprotein-cholesterol (HDL-C) and heart attacks: people with higher levels of HDL-C had lower rates of heart attacks, even after controlling for many confounders. A double-blind randomised controlled experiment involving 12,000 subjects was done from 2012 to 2016 to study a drug that increases HDL-C level. Subjects in the treatment group had higher HDL-C compared to the control group, but the rate of heart attacks were the same in both groups.

(I) The observational studies suggested that increasing HDL-C prevents heart attacks.  
(II) The randomised experiment proved that increasing HDL-C prevents heart attacks.

A. (I) and (II) are true.  
B. Only (I) is true.  
C. Only (II) is true. D. (I) and (II) are false.

4. The Salk vaccine trial in Chapter 1 consisted of two parts: the NFIP study and a randomised experiment. The second part is “randomised” because  
(I) eligible subjects were randomly assigned to treatment and control.  
(II) eligible subjects were randomly chosen from the population.

A. Both (I) and (II).  
B. Only (I).  
C. Only (II).  
D. Neither (I) nor (II).

All 674 employees of a company were asked if they found work stressful. Employees were also classified as old (age 50 years or more) or young (less than 50 years old). The data are summarised in the table. The next two problems are based on this situation.

| Sex | Age group | Stressful | Not stressful | Row sum |
|---|---|---|---|---|
| Female | Old | 53 | 414 | 467 |
| | Young | 11 | 37 | 48 |
| Male | Old | 0 | 16 | 16 |
| | Young | 4 | 139 | 143 |
| Total | Old | 53 | 430 | 483 |
| | Young | 15 | 176 | 191 |

5. For the question of whether age has an effect on work stress, sex is a confounder. This conclusion is justified by
(I) the overall rate of stress among old workers was higher than the overall rate of stress among young workers, i.e., 53/483 > 15/191.
(II) the rate of stress among females, 64/515, is different from the rate of stress among males, 4/159.
A. Both (I) and (II).
B. Only (I).
C. Only (II).
D. Neither (I) nor (II).

6. The statement “The rate of stress among old workers is lower than the rate of stress among young workers.” is true among
(I) female workers. (II) male workers. (III) all workers.
A. (I) only.
B. (II) only.
C. (III) only.
D. (I) and (II) only.
E. (I), (II) and (III).

7. In 2015, 57% of US companies offered health insurance to employees. Among companies that had 100 or more employees, 97% offered health insurance. Among companies employing less than 100 people, the percentage that offered health insurance
(I) can be calculated from the information given.
(II) must be less than 57%.
A. (I) and (II) are true.
B. Only (I) is true.
C. Only (II) is true.
D. (I) and (II) are false.

8. In the Pearson’s father-son data set (shown below), a researcher computes the average height of fathers who are between 65 inches (inclusive) and 66 inches (exclusive), and also the average height of their sons. This yields a single point plotted on a new graph. He proceeds to do the same for other intervals: 66-67, 67-68, etc. as well as 64-65, 63-64, etc., obtaining a new scatter diagram with about 25 points. The correlation of the new scatter diagram is ___ the correlation of the original data set.

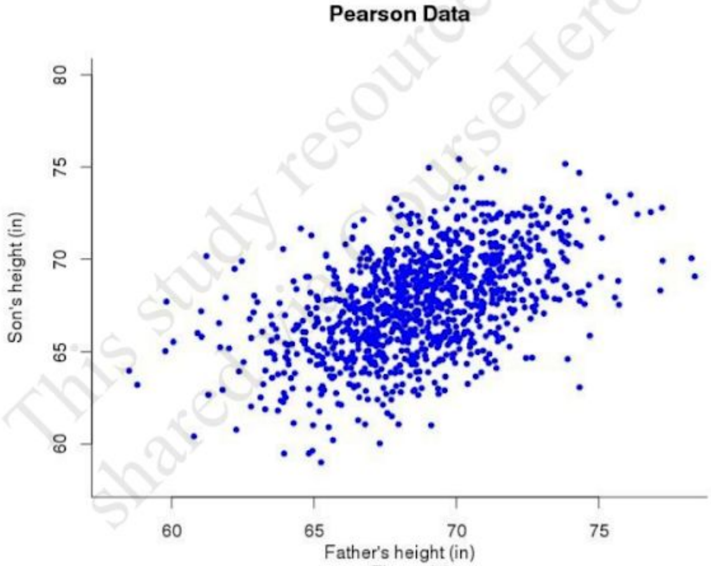


Figure (1)

A. more than
B. similar to
C. less than

9. Students of a university fill out questionnaires giving their year of birth, age (in years), age of father, and so forth. The correlation between student’s age and year of birth is closest to

A. 1
B. 0.5
C. 0
D. -0.5
E. -1

10. From the men in a large country, 13,191 were randomly selected. From the women in the same country, 11,482 were randomly selected. All 24,673 subjects were examined for diabetes. Data are summarised in the table.

| | Diabetes | No diabetes | Row sum |
|---|---|---|---|
| Men | 1251 | 11940 | 13191 |
| Women | 512 | 10970 | 11482 |
| Column sum | 1763 | 22910 | 24673 |

(I) The risk ratio for diabetes of men to women in the country is roughly 2.13.
(II) The odds ratio for diabetes of men to women in the country is roughly 2.24.

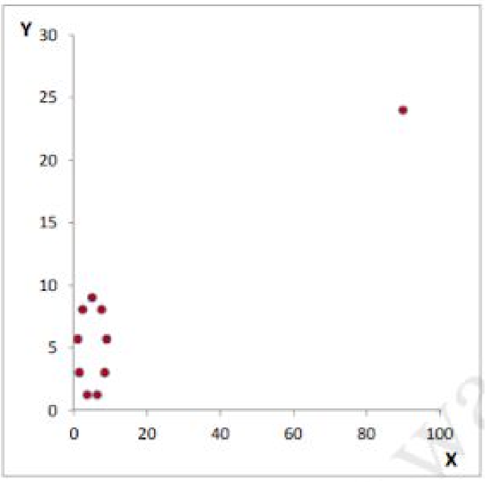
A. (I) and (II) are true.
B. Only (I) is true.
C. Only (II) is true.
D. (I) and (II) are false.

11. As part of their training, air force pilots made two practice landings with their instructors and were rated on their performance. The instructors discussed the performance and ratings with the pilots after each landing. An analysis showed that pilots who made poor landings the first time tended to do better the second time. Conversely, pilots who had good landings the first time tended to do worse the second time.

Choose the most appropriate option among the following:

A. This shows that criticism helps the pilots improve their landings, while praise makes them do worse.
B. The instructors now decided to criticize all first landings, regardless of actual performance. This decision will lead to better ratings on the pilots’ second landing.
C. This can be explained by the effect of regression towards mediocrity.
D. These results are exceptional. Normally, we expect pilots with good ratings on the first landing to do even better the second time, and pilots with poor ratings on their first landing will do worse the second time.

12. The following plot shows an outlier in both the X and Y directions. What will happen if we remove the outlier?

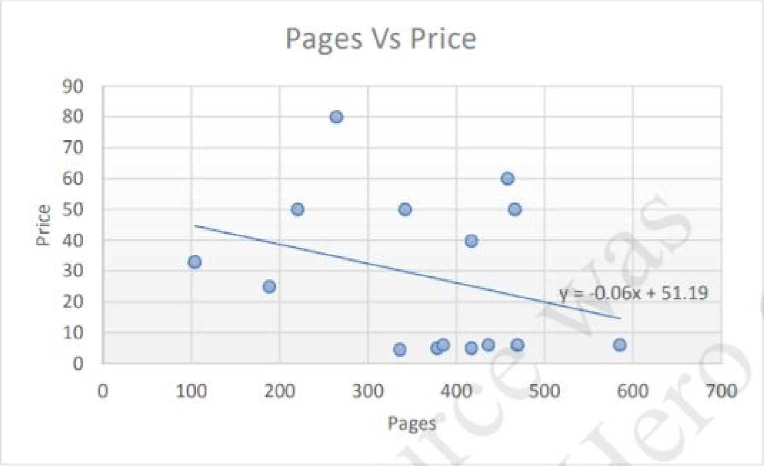


A. The correlation coefficient between X and Y will remain roughly the same.
B. The correlation coefficient between X and Y will decrease.
C. The correlation coefficient between X and Y will increase.
D. The correlation coefficient may increase or decrease, depending on the scales of measurement for X and Y.
E. It is not possible to know what will happen to the correlation coefficient between X and Y.

13. The correlation between height and weight among men age 18-74 is about 0.40.
(I) Taller men tend to be heavier.
(II) If someone eats more and puts on 10 kg, he is likely to get somewhat taller.
A. (I) and (II) are true.
B. Only (I) is true.
C. Only (II) is true.
D. (I) and (II) are false.

14. The following table lists the number of pages, price and the type of 15 books. H means hardcover while S means softcover. The scatter diagram is shown below, together with the regression line of price against page.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Page | 104 | 188 | 220 | 264 | 336 | 342 | 378 | 385 | 417 | 417 | 436 | 458 | 466 | 469 | 585 |
| Price | 32.95 | 24.95 | 49.95 | 79.95 | 4.5 | 49.95 | 4.95 | 5.99 | 4.95 | 39.75 | 5.95 | 60 | 49.95 | 5.99 | 5.95 |
| Type | H | H | H | H | S | H | S | S | S | H | S | H | H | S | S |



(I) For each increase of one page there is an average increase in the price of the book by $0.06.
(II) The regression is used to predict the price of a book of 500 pages. If the book is actually hardcover, the prediction is likely too low.
A. (I) and (II) are true.
B. Only (I) is true.
C. Only (II) is true.
D. (I) and (II) are false.

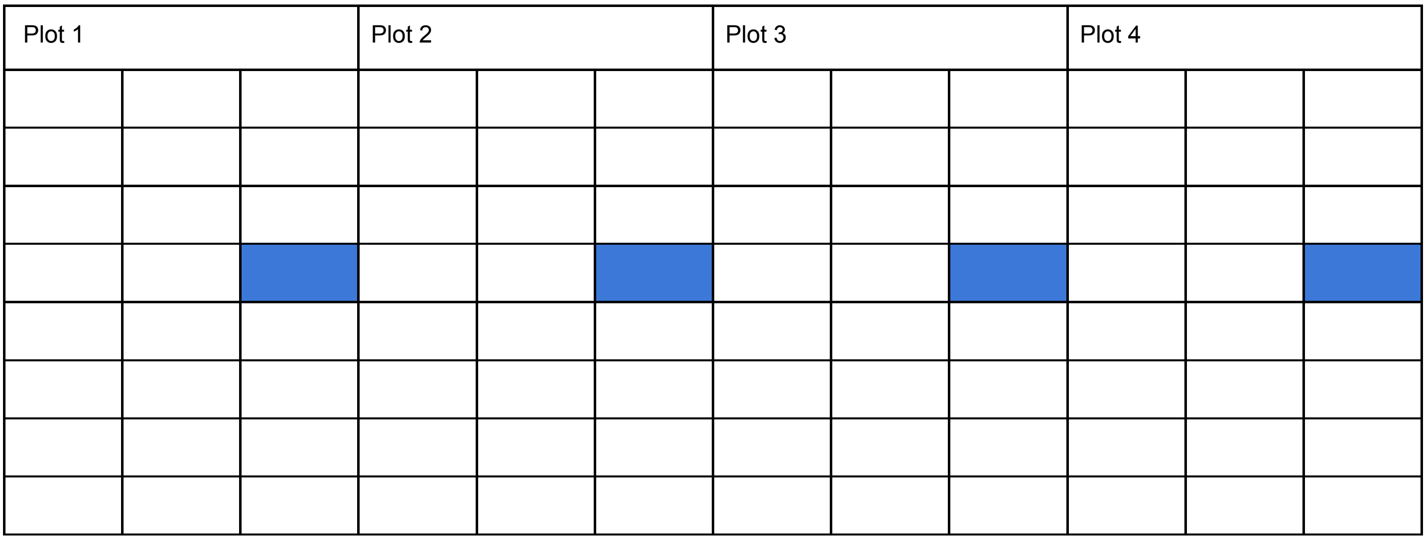
Solutions:
1B. A controlled experiment with randomisation is more reliable than without. In the three randomised experiments, there was no difference in miscarriage rates between control and treatment groups. [Chapter 1 Units 4, 5]
2D. (I) states rate(pessimistic | no diploma) > rate(optimistic | no diploma). This is true, but does not show association. (II): rate(optimistic | no diploma) > rate(optimistic | diploma), and (III): rate(no diploma | optimistic) > rate(no diploma | pessimistic) both show association. [Chapter 1 Unit 6]
3B. (I) is true: this is like how smoking was suspected to cause ill health. (II) is false since the drug failed to reduce heart attacks in the treatment group.
4B. Eligible subjects were not randomly chosen, since parental consent was needed. [Chapter 1 Unit 4]
5D. (I) says age and stress are associated, but nothing about sex. (II) says sex and stress are associated, but this is incomplete: we also need an association between sex and age. [Chapter 1 Unit 6]
6D. (I) is true, since 53/467 < 11/48. (II) is true, since 0/16 < 4/143. But (III) is false: 53/483 > 15/191. This illustrates Simpson’s paradox. [Chapter 1 Unit 9]
7C. It is impossible to know the percentage, but since the percentage in the big companies is larger than 57%, the percentage in the small companies must be less than 57% in order to have an overall percentage of 57%. This is like question 4 in Quiz 1.
8A. This is an ecological correlation. [Chapter 2 Unit 8]
9E. The correlation must be negative: the older students were born earlier. The vast majority of students lie on a straight line, so D can be ruled out.
10A. This is a cohort study, so both RR and OR can be estimated from the data. RR ~ (1251/13191)/(512/11482) = 2.13. The odds for men and women are respectively 1251/11940 and 512/10970, so the ratio is about 2.24. Or we can use the “cross-calculation” to get the answer directly. [Chapter 1 Unit 10, Chapter 2, Unit 3]
11C. This is like taller fathers (good landing on first try) tend to have relatively short sons (poorer second landing). [Chapter 2 Unit 9]
12B. Without the outlier, correlation is almost 0: as we go along the x axis, there is little change in the average y value.
13B. A positive correlation implies (I) in general. But the causal interpretation in (II) doesn’t follow. “Association does not mean causation.” [Chapter 2 Unit 3]
14C. (I) is false, since the slope should be negative. (II) is true since the hardcover books are more expensive than softcover books, and the regression line tries to fit in between.

1. A gourmet food magazine wants to know how its readers feel about serving beer with various types of food. Note that not all readers are subscribers of the magazine. The magazine sends surveys to 1,000 randomly selected names from its list of subscribers. Which of the difficulties in sampling is the magazine most likely to face?

(i) Using the wrong sampling frame.
(ii) A low response rate.

(A) (i) only
(B) (ii) only
[✓ (C) (i) and (ii)]
(D) None of the above

2. Suppose a large piece of land is divided into several plots. Each plot is further divided into several smaller subplots. A sample of subplots is to be selected in the following way:
(a) A subplot is randomly selected from one particular plot;
(b) For the remaining plots the subplot that is in the same position as the selected subplot in (a) is selected.
For example, in the diagram below, say the shaded subplot in Plot 1 is randomly selected, then the shaded subplots in the remaining plots are also selected.



This sampling scheme is called
(A) Simple random sampling
(B) Stratified sampling
[✓ (C) Systematic sampling]
(D) None of the above

3. Which of the following are advantages of a case-control study over a randomized Experiment?

(i) It saves time and money.
(ii) It makes establishing causal relationship between the exposure and response variables easier.
(iii) It presents fewer ethical issues.

(A) (i) and (ii) are correct.
[✓ (B) (i) and (iii) are correct.]
(C) (ii) and (iii) are correct.
(D) All are correct.

4. Without random sampling, which of the following can happen?

(i) The results may not be extend to a larger population.
(ii) The researcher will have a much easier time getting participants for the study, resulting in a larger sample size, and more accurate data.

[✓ (A) (i) only]
(B) (ii) only
(C) (i) and (ii)
(D) None of the above

5. The mid-term test of a certain module has a high (30%) failure rate. The instructor of the module conducted a few special sessions for the entire class after the test. He then randomly selected ten students to sit for another test (with similar scope and level of difficulty). Nine of these students passed the second test. To determine whether the special sessions are effective (in helping students to pass the test), the instructor carried out a hypothesis test. Which of the following is correct?

(A) At 5% level of statistical significance, the instructor is certain that the special sessions will reduce the failure rate.
(B) At 5% level of statistical significance, the instructor is certain that the special sessions will not reduce the failure rate.
(C) The probability that the null hypothesis is true is about 0.15.
(D) None of the above.

6. A certain infectious disease is spread only by direct contact. Suppose the probability for a healthy person to get infected by a single direct contact is 0.02. What is the chance that a healthy person gets infected with the disease by the fourth time he comes into contact with an infected person? (You may assume the outcome of coming into contact with an infected person is independent from each other.)
(A) 0.0776
(B) 0.0188
(C) 0.08
(D) None of the above.

7. An insurance company sells a certain health insurance policy that has a maximum payout of $30,000 when a customer makes a claim. Suppose that, each year, 2% of the customers receive the full claim of $30,000, 3% receive a claim of $20,000, not more than 5% receive a claim of $10,000, while the rest of the customers do not submit a claim.
In order not to make a loss for sure, the minimum amount that the insurance company should charge per year for this policy is
(A) below $1,500.
(B) between $1,500 and $1,800.
(C) between $1,800 and $2,000.
(D) more than $2,000.

8. Among 100,000 women with negative mammograms (screening for breast diseases), 20 will be diagnosed with breast cancer in 2 years, whereas 1 woman in 10 with positive mammograms will be diagnosed with breast cancer in 2 years. Suppose that 10% of the general population of women will have a positive mammogram. What is the probability that a woman who develops breast cancer over the next 2 years has a negative mammogram?
(A) 0.0177
(B) 0.0196
(C) 0.00018
(D) 0.0002

9. For any two events A and B with both P(A) and P(B) > 0, the conditional probability P(B|A)
(A) is greater than P(B) if A and B are independent events.
(B) is equal to P(B) if A and B are mutually exclusive events.
(C) can be greater than or less than P(B).
(D) None of the above.

10. The organizer of a running event says that the run will be cancelled in the event that either (i) it rains on the run day; or (ii) the response for the run is poor.
Consider the following statements:

> (I) The probability that the run is cancelled is the sum of the probability that it rains on the run day and the probability that the response for the run is poor.
> (II) The probability that it rains on the run day and the run is cancelled is equal to the probability that it rains on the run day.

(A) Both (I) and (II) are correct.
(B) Only (I) is correct.
(C) Only (II) is correct.
(D) Both (I) and (II) are incorrect.

11. Suppose $u$ and $v$ are two adjacent vertices in a network whose order is 50. If the degree centrality measures of $u$ and $v$ are 1/49 and 2/49 respectively, which of the following statements is true?
(A) Bcen($v$) is twice Bcen($u$).
(B) Bcen($u$) is twice Bcen($v$).
(C) Ccen($v$) will always be greater than Ccen($u$).
(D) Ccen($u$) will always be greater than Ccen($v$).

12. A network has the following adjacency matrix:

| | $u$ | $v$1 | $v$2 | $v$3 | $v$4 | $v$5 |
|---|---|---|---|---|---|---|
| $u$ | 0 | 1 | 1 | 1 | 0 | 0 |
| $v$1 | 1 | 0 | 0 | 0 | 0 | 0 |
| $v$2 | 1 | 0 | 0 | 0 | 1 | 0 |
| $v$3 | 1 | 0 | 0 | 0 | 1 | 0 |
| $v$4 | 0 | 0 | 1 | 1 | 0 | 1 |
| $v$5 | 0 | 0 | 0 | 0 | 1 | 0 |

Using the definition of closeness centrality measure of a vertex $u$, namely,

$$Ccen(u) = \frac{\sum_{i=1}^{n-1} d(u, v_i)}{n-1},$$

what is the value of Ccen($u$)?

(A) 9/5

(B) 2

(C) 11/5

(D) 8/5

13. Recall that in our discussion of the movie network in Chapter 5, two actors are adjacent in the graph if they have featured in a movie together. The Bacon number of an actor is the distance of this actor's vertex in the network from the Kevin Bacon vertex. Suppose that the Bacon number of an actor A was initially computed to be 3. However, it was later discovered that there were two errors in the database. They are:

(i) There is a movie, featuring Kevin Bacon but not actor A, that was omitted from the database.

(ii) There is a movie which was initially thought to have featured actor A but actually did not feature him.

After correcting these errors, which of the following statements is correct?

(A) The Bacon number of actor A is now 1.

(B) The Bacon number of actor A now could be 2, 3 or bigger than 3.

(C) The Bacon number of actor A is now 2.

(D) The Bacon number of actor A is still 3.

For Questions 14 and 15, consider the following:

In addition to the three centrality measures introduced in the module, there are many other centrality measures proposed by researchers. For a vertex $u$ in the network, let Acen($u$) be one such other centrality measure.

Suppose we have a network of order 500 and the centrality measures Acen($u$), Bcen($u$) (betweenness centrality measure) and Ccen($u$) (closeness centrality measure) were computed for each vertex $u$.

14. While investigating the linear association between measures Acen and Bcen using a scatter plot, it was found that miraculously, the correlation coefficient r is equal to 1. If $v$ is a vertex in the network with Acen($v$) = 0.56, what can be said of the value Bcen($v$)?

(A) It is smaller than 0.56.

(B) It is 0.56.

(C) It is larger than 0.56.

(D) It cannot be determined with the information given.

15. Suppose the 500 vertices in the network represent 300 teenagers (aged 16-20) and 200 adults (aged 21 and above). It was discovered that Acen and Ccen are negatively correlated among the teenagers but positively correlated among the adults.

When the data for teenagers and adults are combined, the correlation coefficient between Acen and Ccen

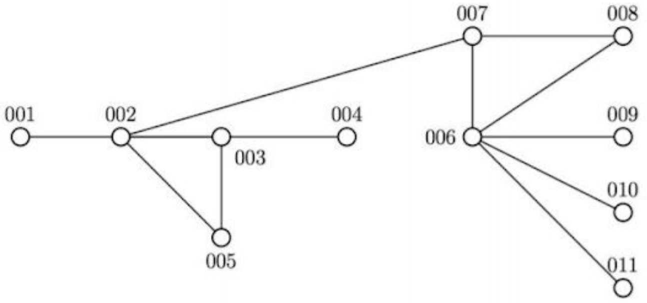
(A) cannot be determined with the information given.

(B) will be positive.

(C) will be negative.

(D) will be zero.

16. The following diagram represents a network of undercover agents that you are overseeing. Each vertex represents an individual and an edge between two vertices represents the existence of a direct communication channel between the two individuals. You may assume that in disseminating information to all the agents, you would always keep the number of communications between individuals to be as small as possible, in order to minimize the risk of exposing the agents' identities.



Your enemy intends to convert one of your agents in the network to working for him and the job of this converted spy is to intercept as many messages and as much information that is being passed around as possible. The degree, closeness and betweenness centrality measures for all the vertices, as defined in the module GER1000, are given in the table below:

| Vertex | Dcen | Ccen | Bcen | Vertex | Dcen | Ccen | Bcen |
|---|---|---|---|---|---|---|---|
| 001 | 0.1 | 2.8 | 0 | 007 | 0.3 | 1.8 | 0.56 |
| 002 | 0.4 | 1.9 | 0.6 | 008 | 0.2 | 2.3 | 0 |
| 003 | 0.3 | 2.5 | 0.2 | 009 | 0.1 | 2.9 | 0 |
| 004 | 0.1 | 3.4 | 0 | 010 | 0.1 | 2.9 | 0 |
| 005 | 0.2 | 2.6 | 0 | 011 | 0.1 | 2.9 | 0 |
| 006 | 0.5 | 2.0 | 0.53 | | | | |

Which vertex would your enemy most likely try to convert?
(A) 006 since it has the largest degree centrality measure.
(B) 007 since it has the smallest closeness centrality measure.
(C) 002 since it has the largest betweenness centrality measure.
(D) 004 since it is in the centre of the network.

Some Definitions
Let x and y be two vertices in a network (or graph). There can be several paths connecting x and y, including a path with the smallest number of edges. The distance between x and y, d(x,y), is defined as the number of edges in this path.
Let u be a vertex in a network (or graph) of n vertices. The degree centrality measure of u is the number of edges attached to it, divided by n-1. The closeness centrality measure is the sum of the distances between u and all other vertices, divided by n-1.

1. C
2. C
3. B
4. A
5. D
6. A
7. B
8. A
9. C
10. C
11. D
12. D
13. B
14. D
15. A
16. C

1. Which of the following statements is correct:
a. An observational study does not have a control group.
b. A case-control study cannot be a controlled experiment.
c. We cannot have randomness in observational studies.

2. A study on a new invented drug is being tested on rabbits. The rabbits being assigned to the treatment group will receive the dug while those in the control group will not. Due to limited supply, the researchers can only have 2 rabbits per day. Thus they decided to send the rabbits coming on odd days to treatment group and those coming on even days to control group.
a. This is a cohort study.
b. This is a case-control study.
c. This is a randomized controlled experiment.
d. This is a non-randomized controlled experiment

3. A study on “Do healthy sleep habits lead to greater happiness?” is conducted in two countries, UK and US. The researcher used a survey to collect information from participants.
The survey asked the participants to rate they sleep habits using very unhealthy, unhealthy, healthy, and very healthy. However, the researcher decided to merge the four categories into two. Very unhealthy and unhealthy were then considered as unhealthy. Very healthy and healthy were then considered as healthy. The data shows that people with better sleep habits are more likely to be happy.
Which of the following statements is correct?
a. There is likely to be an ecological fallacy in the study, since it was done on two separate populations.
b. We have a chance to observe a Simpson’s paradox, when comparing the original data which has four categories of sleep habits with the merged data which has only two categories of sleep habits.
c. If the study was done on a voluntary sample, we cannot conclude that there is a possible association between healthiness of sleep habits and happiness.
d. None of the above

4. A box contains 3 marbles: Red, Blue and Green. Randomly pick one marble from the box, and then, without replacement, pick another one from the box. List out all possible outcomes. And calculate the probability of the event that, the second marble picked out is a red one.

5. Given that the probability of raining tomorrow is 0.5, and the probability of raining the day after tomorrow is 0.8. Using the information given, can we determine the probabilities of the following events?
Raining tomorrow and the day after tomorrow.
Raining tomorrow or the day after tomorrow.

6. Given that the probability that Tom will eat dinner today is 0.5, and the probability that Tom will eat supper is 0.6. Moreover, the probability that Tom will eat both dinner and supper is 0.3. Are the two events independent?
Tom will eat dinner
Tom will eat supper

7. Given that the probability that Jerry will eat dinner but not supper today is 0.5, and the probability that Jerry will eat supper but not dinner is 0.6. Moreover, the probability that Jerry will eat both dinner and supper is 0.3. Are the two events independent?
Jerry will eat dinner but not supper
Jerry will eat supper but not dinner

8. Two fair dice are rolled. Let X be the outcome of the first dice and Y be the second. Compute the expected value of X*Y.

9. A drug is invented for some disease, which has a fatal rate of 0.6. Conduct hypothesis test, using a critical value of 0.05, to test whether the drug is effective given the following situations:
a. 4 out of 5 patients survived
b. 3 out of 5 patients survived

10. Two fair dice are rolled. What is the conditional probability that one dice lands on 6, given that the dice land on different numbers?

~~11~~. Consider the following adjacency matrix. What is the distance between vertices a and b.

| | a | b | c | d |
|---|---|---|---|---|
| a | 0 | 1 | 0 | 1 |
| b | 1 | 0 | 1 | 1 |
| c | 0 | 1 | 0 | 1 |
| d | 1 | 1 | 1 | 0 |

ANS:

1. B

2. D

3. D

4. {R,B};{B,G};{R,G};{B,R};{G,B};{G,R}  P(R in the second draw) = ⅓

5. We cannot, unless independence and mutual exclusivity are known. However, the weather of consecutive days are likely to be dependent. And apparently, “It will be raining tmw” and “it will be raining the day after tmw” are not mutually exclusive.

6. Yes. Though eating dinner may affect Tom’s desire for a supper, the probabilities given satisfies the definition of independence.

7. No, these two events are mutually exclusive, thus they cannot be independent.

8. 12.25

9. Hypothesis: The drug is not effective.
Case (a): P(4 patients survived) = $(0.4)^4(0.6)$ 5 = 0.0768
P(5 patients survived) = $(0.4)^5$ = 0.01024
Thus, P-value = 0.08704
Case (b): P(3 patients survived) = $(0.4)^3(0.6)^2$ 10 = 0.2304
Thus P-value = 0.31744
In either case, we cannot reject the hypothesis, thus the drug might be ineffective.
If you don’t know combinatorics, you can ignore part (b). The combinatorics formulas will not be tested, thus such cases will not appear in the exam.

10. ⅓

11. The distance is 1 because the two vertices are adjacent.