**GER1000 2020 Sem 1**

**Quiz 6 and solutions**

A researcher wishes to examine the impact that smoking has on the risk of getting lung cancer in Town *X*. Town *X* has a population of 500,000 people, of which 5,780 people have lung cancer, whilst the rest do not. He randomly sampled 10% of the people with lung cancer and 20% of the people without lung cancer and enquires about their smoking history. From here we assume the following

* The number of people with lung cancer is accurate and known to the researcher.
* All the sampled participants agreed to participate in the researcher’s study.
* All participants gave truthful responses regarding their smoking history.

The results are summarized in the contingency table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Lung cancer | No lung cancer | Row total |
| Smokers | 457 | 20,246 | 20,703 |
| Non-smokers | 121 | 78,598 | 78,719 |
| Column total | 578 | 98,844 | 99,422 |

Use the information above to answer the following 2 questions:

1) What is the sample risk ratio for lung cancer between smokers and non-smokers, correct to 1 decimal place?

(A) 14.4

(B) 14.7

(C) The sample risk ratio cannot be calculated in this scenario as this is a cohort study.

(D) The sample risk ratio cannot be calculated in this scenario as this is a case-control study.

*Explanation: The sample risk ratio can always be calculated regardless of the type of observational study. In this case, Similar to Chapter 5, Unit 1, Slide 5.*

2) Based on the above scenario, which of the following statements is/are true?

(I) The sample risk ratio of lung cancer between smokers and non-smokers does not give a good estimate of the population risk ratio.

(II) The sample odds ratio of lung cancer between smokers and non-smokers gives a good estimate of the population odds ratio.

(A) (I) only

(B) (II) only

(C) Both (I) and (II)

(D) Neither (I) nor (II)

*Explanation: This is a case-control study as samples are taken from the disease groups; furthermore, the population risk ratio cannot be accurately estimated from a case-control study in general. Refer to Chapter 5, Unit 2. Regardless of cohort or case-control studies, the sample odds ratio will give a good estimate of the population odds ratio.*

3) Which one of the following statements is/are true regarding cohort and case-control studies?

(I) Case-control studies are controlled experiments, whilst cohort studies are observational studies.

(II) Case-control studies are better for rare diseases, compared to cohort studies.

(A) (I) only

(B) (II) only

(C) Both (I) and (II)

(D) Neither (I) nor (II)

*Explanation: Both case-control and cohort studies belong within the category of observational studies. The “control” in the term case-control does not refer to controlled experiments. Therefore (I) is false. For (II) refer to Chapter 5 Unit 1 Slide 10.*

4) Given that the odds of an event A occurring is 5, what is the risk of event A happening?

(A) 5

(B) 1/5

(C) 1/6

(D) 5/6

*Explanation: Recall that odds(A)* = . *Solving for this equation with odds(A) = 5, this will give us risk(A) = 5/6.*

5) A certain town has a population of 100,000, of which 20,000 are smokers and 80,000 are non-smokers. It is known that for the population, r(lung cancer | smoker) = 5% and r(lung cancer | non-smokers) = 1%. Suppose that 50% of lung cancer patients are chosen by simple random sampling, while 10% of residents without lung cancer are chosen by simple random sampling. By calculating r(lung cancer | smoker) in the **sample** using expected numbers,what can we say about

U = sample r(lung cancer | smoker) and V = population r(lung cancer | smoker)?

(A) U > V

(B) U = V

(C) U < V

*Explanation : Refer to Chapter 4 Unit 1 Slides 4-6 to see how expected numbers are calculated. Within the sample, there will be a total of 2400 smokers, of which 500 have lung cancer. Therefore, sample r(lung cancer | smoker) = 5/24 = 21%, which is more than 5%, the population r(lung cancer | smoker).*

6) Suppose A and B are events with P(A) = 0.4 and P(B) = 0.7. Which of the following statements is/are correct?

A and B can be mutually exclusive.

P(A and B) = 0.4 + 0.7.

1. Both (I) and (II)
2. Only (I)
3. Only (II)
4. Neither (I) nor (II)

*Explanation: It is impossible for the probability of any event occurring, be it single or composite, to exceed 1. We can see that statement (II) is incorrect, since the statement asserts that P(A) + P(B) = 1.1 > 1. For statement (I), even if A and B are mutually exclusive events, then P(A or B happens) = P(A) + P(B) = 0.4 + 0.7 = 1.1, which still cannot hold for the same reason. Hence (I) is also incorrect. See Chapter 5, Unit 1, Slide 4 and Unit 2, Slide 6.*

7) I have a fair 12-sided (dodecahedron) die with sides labelled 1, 2, …, 12 respectively. See <https://images.app.goo.gl/FGseg31kiTAW7psK6> for an example of a fair 12-sided die. I also have a fair 6-sided die with sides labelled 1, 2, …, 6 respectively. I roll the first die on a table with a standard protocol, then roll the second die on another table also with a standard protocol. What is the probability that the sum of the numbers appearing face up on the two dice is 11?

1. 1/12
2. 5/36
3. 1/18
4. 1/9

*Explanation: One way to obtain a sum of 11 is to get 10 from the first die and 1 from the second die, which can be written as (10,1). Since the dice are fair, and their rolls are independent, the probability of (10,1) is (1/12) x (1/6). The only other ways to get a sum of 11 are (9,2), (8,3), (7,4), (6,5) and (5,6), each of which has probability (1/12) x (1/6. Since the 6 possibilities are mutually exclusive, we sum their probabilities. Hence P(sum is 11) = 6 x (1/12) x (1/6) = 1/12. See Chapter 5, Unit 2, Slides 5, 6, 10 and 11.*

8) Tom wants to study the association between the intelligence quotient (IQ) and emotional quotient (EQ) of university students in Singapore. From each faculty in NUS, he gathered the average IQ and average EQ scores of students, and plotted the average IQ scores of students against the average EQ scores of students across different faculties. A correlation coefficient of 0.75 was observed. Without any further investigation, he concludes that “the correlation coefficient between the average IQ and the average EQ scores of students across all the different faculties in Singaporean universities is 0.75”. Assuming that all data collected on IQ and EQ scores were fully accurate, which of the following statements is true?

1. Tom committed ecological fallacy
2. Tom committed atomistic fallacy
3. None of the other options

*Explanation: If Tom used this correlation between average IQ scores and average EQ scores of students to conclude that the correlation coefficient between IQ and EQ scores for individual students in NUS is 0.75, then he would have committed the ecological fallacy. In addition, since Tom did not use any correlation based on individuals to conclude the correlation on the aggregate level, he did not commit the atomistic fallacy as well. See Chapter 3, Unit 6, slides 72 and 73. Here, the mistake Tom made was a simple generalisation of the result obtained for NUS faculties to all faculties in all Singaporean universities, since the faculties in NUS may not be representative of all faculties in Singaporean universities.*

9) Suppose I wish to find the average intelligence quotient (IQ) of all Primary 5 children studying in local schools in Singapore. I first selected a simple random sample of 10 schools out of all local primary schools in Singapore. Then I asked all the Primary 5 children in these chosen 10 schools to take an IQ test. Finally, I obtained the average value of all the IQ scores of children who took the test, which was 106. Which of the following statements is/are correct?

1. The parameter in this study is the average IQ of all Primary 5 children who took the IQ test.
2. Stratified sampling was employed in this study.

(A) Both (I) and (II)

(B) Only (I)

(C) Only (II)

(D) Neither (I) nor (II)

*Explanation: The parameter in this study is the average IQ of all Primary 5 children studying in local schools in Singapore. 106 is only a sample estimate of the actual parameter. Hence statement (I) is incorrect. In stratified sampling, the population is divided into groups (strata) and then we randomly obtain a sample from each group. In cluster sampling, the population is first divided into groups (clusters). Then we take a random selection of clusters from all clusters, and include all units in the chosen clusters to comprise our sample. Here, cluster sampling is employed, where each school is a cluster. Hence statement (II) is also incorrect. See Chapter 3, Unit 3, Slides 31-32, 34-36, and Unit 5, Slide 60.*