

## GER1000 Exam Practice Questions

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.

Ans: D. (I) Outliers should be handled with care. (II) Depends. Outliers can either increase, or decrease the strength of the correlation.

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

Ans: B.

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.

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. 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 is 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.

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. 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)

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. 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

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. 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

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

$P(\text{drug user} \mid \text{tested positive}) = 4,750/9,500 = \frac{1}{2}$ .

[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. 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

Ans: D. Expected profit =  $\$800 * 97\% + -\$4200 * 1.5\% + \$ (800-20,000) * 1.5\% = \$425$ .

9. A policeman in the U.S. 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).~~

(II) Being convicted of robbery is independent of ethnicity.

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

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. 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 '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.

Ans: B. The null hypothesis should be that the dice is fair, or not biased.

p-value =  $P(\text{dice lands 4 or more times on a 7, given that the dice is not biased})$

=  $P(\text{dice lands 4 times on 7, given that the dice is fair}) + P(\text{dice lands 5 times, given that...})$

=  $5 \times (1/20)^4 \times (19/20) + (1/20)^5 = 0.00003 < 0.01$ .

(The p-value is the probability that 4 **or more** '7's are observed, **if** the dice is fair).

11. 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?

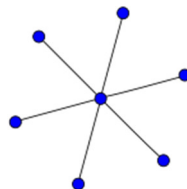
- (I) Joy is a resident in this country. The probability that she has blood type A is 0.41.
- (II) A person is selected at random from the resident population. The probability that he/she has blood type A is 0.41.

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

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. 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)

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. 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?

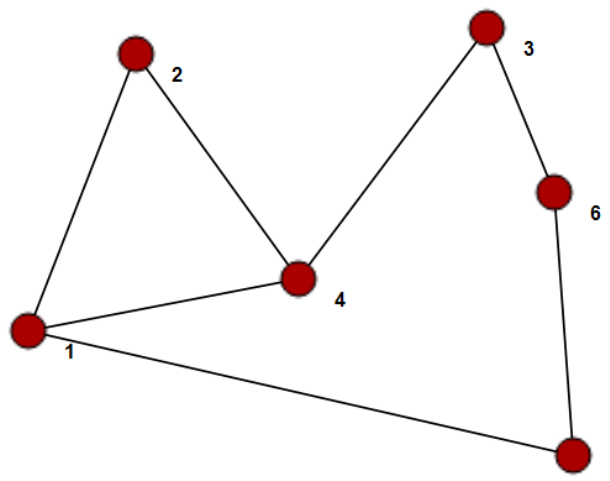
- A. 1,1,0,1,0
- B. 0,1,1,0,1
- C. 0,1,0,0,1
- D. 1,1,1,0,1

Ans: C. Note that the adjacency matrix is symmetric along the diagonal. For example, the value in the 1<sup>st</sup> row, 3<sup>rd</sup> column should be the same as the value in the 3<sup>rd</sup> row, 1<sup>st</sup> column. (If vertex 1 and 3 are adjacent, this also means that vertex 3 and 1 are adjacent.)

Also, the values along the diagonal (1<sup>st</sup> row, 1<sup>st</sup> column; 2<sup>nd</sup> row, 2<sup>nd</sup> 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. 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

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. 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).

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. 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.

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  $9 \times 10 = 90$  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. Recall that the Old-Age Support Ratio (OASR) for a population is defined as

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

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):

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

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.

Ans: B. The numerator for NSR contains a larger group, compared to the numerator for OASR.

18. You have been engaged by the government of Pulau Guleam to project the number of 75-year-olds 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 the **age-specific** 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
<b>Population</b>	11879	10726	9572	8944	8317	7971	7343	6716	6146	5577
<b>Death rate</b>	21.2	23.5	25.8	29.2	32.7	36.3	39.7	43.2	48.2	53.2
<b>Migration rate</b>	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. 8527 (was previously 8543)
- B. 8422
- C. 8748
- D. 8007
- E. Cannot be determined, as the birth rates are not given.

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. 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 6.974 (4.258).
- C. It cannot be calculated or estimated. We need to know the ASFRs for all childbearing ages from 15 to 49 years.

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. The Sex Ratio at birth for a certain country was 1070 for 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?

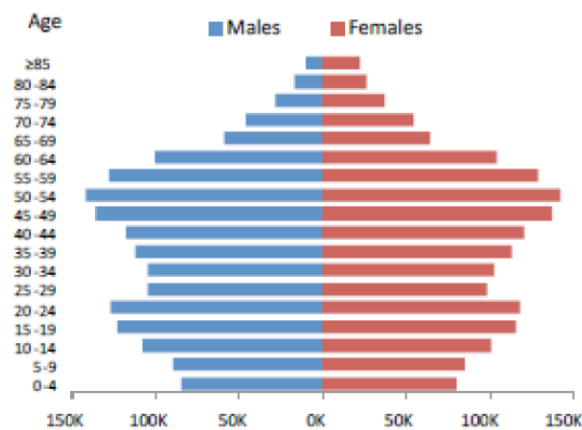
- A. 1017
- B. 965
- C. We cannot calculate its exact value, but it is less than 1070
- D. It should still be 1070

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  $95\% \times 1070/2070 = 1016.5/2070$ . 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. 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.

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.

END