Exercise 4

1. There are 5 red balls and 5 blue balls in a bag. Mary picks two balls from the bag, one after another without replacement. She wins a prize if the two balls she picks are of the same colour. The first ball she picks is red. What is the probability that Mary wins a prize?

   (A) $\frac{2}{9}$
   (B) $\frac{1}{2}$
   (C) $\frac{4}{9}$
   (D) $\frac{1}{9}$

   Answer is (C). The answer is simply the conditional probability

   $$P(\text{Second ball is red} \mid \text{First ball is red}) = \frac{\frac{5}{10} \times \frac{4}{9}}{\frac{5}{10}} = \frac{4}{9}.$$

   Note that the answer is simply the probability of choosing 1 red ball, out of the remaining 9 balls in the bag. This is because we have already fixed the event of first ball being red, so the conditional probability reduces to $\frac{4}{9}$ when picking the second ball.

2. A glass-making machine produces tempered glass for mobile phones. The tempered glass produced are either grade A (good), grade B (minor defects) or grade C (unacceptable). The glass produced are put through an inspection machine that is able to detect any grade C glass and discard it. We may assume that the inspection machine makes no mistakes in the detection process, meaning that given a piece of grade C glass, the probability of it being discarded is 1. At the same time, given a piece of grade A or B glass, the probability of it being discarded is 0.

   Suppose the glass-making machine produces grade A glass 90% of the time, grade B glass 2% of the time and grade C glass 8% of the time. If a piece of glass is not discarded, what is the probability (computed to 3 significant figures) that it is grade A?

   (A) 0.900
   (B) 0.920
   (C) 0.978
   (D) There is not enough information for the probability to be computed.

   Answer is (C). We want to compute the probability

   $$P(\text{Grade A} \mid \text{Not grade C}).$$

   By conditional probability,

   $$P(\text{Grade A} \mid \text{Not grade C}) = \frac{P(\text{Grade A and Not grade C})}{P(\text{Not grade C})} = \frac{P(\text{Grade A})}{P(\text{Grade A or B})} = \frac{0.9}{0.92} = 0.978.$$

   Note that a piece of glass that is (Grade A and Not grade C) is certainly (Grade A). Similarly, a piece of glass that is (Not grade C) is certainly (Grade A or B).

3. Which of the following statements is/are true of normal distributions? Select all that apply.

   (A) The density curve of a normal distribution is symmetrical about its mode.
   (B) $N(1, 4)$ has a standard deviation of 2.
   (C) A normal distribution can be left-skewed.
   (D) A normal distribution can be right-skewed.

(A) and (B) are correct. Since the density curve of any normal distribution is symmetrical about its mean, and its mean is equal to its mode, it is symmetrical about its mode. A symmetrical distribution cannot be left-skewed or right-skewed. Lastly, since $N(1, 4)$ has variance 4, it has standard deviation equal to $\sqrt{4} = 2$.

4. A study was conducted on 1000 students in a secondary school on 11th January 2021. On that day, of these 1000 students, 80% of the students took the MRT (mass rapid transit) to school. Among those students that took the MRT to school, 10% were late for school. Among those students who did not take the MRT to school, 20% were late for school.

   Among those students who participated in the study and were late for school, a student was randomly chosen, with every student having the same chance of being selected. What is the probability (correct to 1 decimal place) that the chosen student did not take the MRT?

   (A) 20.0%

   (B) 10.0%

   (C) 33.3%

   (D) 30.0%

   Answer is (C). The total number of students is 1000 and since $P(\text{Take MRT}) = 0.8$ and $P(\text{Did not take MRT}) = 0.2$, hence 800 students took MRT and 200 students did not.

   Among 800 students who took MRT to school, 10% were late. That is,

   $$P(\text{Late} \mid \text{Take MRT}) = 0.1,$$

   so there were 80 students who took the MRT and were late.

   Among 200 students who did not take MRT to school, 20% were late. So

   $$P(\text{Late} \mid \text{Did not take MRT}) = 0.2,$$

   so there were 40 students who did not take the MRT and were late.

   Put the calculated figures in a 2×2 table as shown below.

	Late	Not late	Row total
Took MRT	80	720	800
Did not take MRT	40	160	200
Column total	120	880	1000

   From the table, we can calculate

   $$P(\text{Did not take MRT} \mid \text{Late}) = \frac{40}{120} = 33.3\% \quad \text{(1 decimal place)}.$$

5. A game is played using a fair six-sided die, a pawn and a simple board as shown below. (A pawn is a chess piece.)

S	1	2	3	4	5	E

   Initially, the pawn is placed on square S. The game is played by throwing the die and moving the pawn back and forth in the following manner:

   S 1 2 3 4 5 E 5 4 3 2 1 2 3 .....

Thus, for example if the first and second throws of the die give a "5" and "4" respectively, the final position of the pawn will be on square "3", because the first throw would send the pawn to square 5, and the second throw would then send the pawn from square "5" to square "3".

The game will stop only when the pawn stops at square "E" after a die roll, passing by "E" **does not** end the game.

Let X denote the number of throws of the die required to move the pawn such that it stops at square "E". Which of the following statements is/are true?

(I) $P(X = 2) = \frac{5}{36}$.

(II) The events $X = 1$ and $X = 2$ are mutually exclusive.

(A) Only (I).

(B) Only (II).

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

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

Answer is (D). The event $X = 2$ can only happen when we do not roll a 6 on the first roll, but our first two rolls add up to 6. Note that it is impossible to hit "E" a second time in two die rolls because it is 16 moves away from the starting point. To win in 2 rolls, we need one of the following outcomes to happen:

$$(1, 5), (2, 4), (3, 3), (4, 2), (5, 1).$$

Here, (x, y) denotes getting x on the first roll and y on the second roll. Each of these outcomes has a probability of $\frac{1}{36}$, so $P(X = 2) = \frac{5}{36}$ and statement (I) is correct. Statement (II) is also correct because if we only need one throw to end the game, then we must roll 6 on the single throw, which is mutually exclusive from the above 5 described outcomes corresponding to $X = 2$.

6. Tom selects a child at random from a population of children. Let

   - A be the event a child of age < 3 is selected;
   - B be the event a child of age < 5 is selected.

   It is known that $P(A) > 0$. Which of the following must be true?

   (I) $P(A \text{ or } B) < P(A) + P(B)$.

   (II) $P(A) \leq P(B)$.

   (A) Only (I).

   (B) Only (II).

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

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

   Answer is (D). Every child of age < 3 is also of age < 5. Thus, event B must occur if event A occurs. This means that $P(A) \leq P(B)$, that is statement (II) is true.

   Furthermore, since whenever event A occurs, event B must occur, we have $P(A \text{ or } B) = P(B)$. So comparing $P(A \text{ or } B) = P(B)$ and $P(A) + P(B)$, we conclude that $P(A \text{ or } B) < P(A) + P(B)$ since $P(A)$ is strictly bigger than 0. Thus, statement (I) is true.

7. Benny is a messy student who keeps all his coloured socks in a box. The box contains a total of 4 blue and 2 yellow socks. While running late for class, he randomly selects (without replacement) two socks out of the box to wear before leaving the house. Assume the socks are indistinguishable from one another in all respects other than their color.

   What is the probability that Benny will end up wearing a pair of matching coloured socks when leaving the house?

(A) $\frac{1}{3}$.

(B) $\frac{2}{5}$.

(C) $\frac{7}{18}$.

(D) $\frac{7}{15}$.

Answer is (D). To get a pair of matching socks, Benny needs to get either 2 blue or 2 yellow socks.

$$\begin{aligned} P(\text{2 blue socks}) &= \frac{4}{6} \times \frac{3}{5} = \frac{12}{30} \\ P(\text{2 yellow socks}) &= \frac{2}{6} \times \frac{1}{5} = \frac{2}{30} \end{aligned}$$

Since the events (2 blue socks) and (2 yellow socks) are mutually exclusive, the probability of getting a pair of matching socks is simply the sum $\frac{12}{30} + \frac{2}{30} = \frac{7}{15}$.

8. A standard deck of 52 playing cards comprises of 4 suits (Clubs, Diamonds, Hearts, Spades), each suit with 13 cards of distinct ranks (A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K).

   Let P_1 be the probability that a card randomly selected from the deck is a "2".

   Let P_2 be the probability that a card randomly selected from the deck is a "2", **given that it is a "Spade"**.

   Which of the following statements is correct?

   (A) $P_1 = P_2$.

   (B) $P_1 < P_2$.

   (C) $P_1 > P_2$.

   Answer is (A). To calculate P_1, since there are four "2"'s in the deck, one of each suit, so $P_1 = \frac{4}{52} = \frac{1}{13}$. To calculate P_2, let A be the event that the card selected is a "2", and B be the event that the card selected is a "Spade". Then

$$\begin{aligned} P_2 = P(A \mid B) &= \frac{P(A \text{ and } B)}{P(B)} \\ &= \frac{\frac{1}{52}}{\frac{13}{52}} \\ &= \frac{1}{13}. \end{aligned}$$

   Thus, $P_1 = P_2$.

9. Systematic sampling with interval length $k = 20$ is to be utilised for an exit poll at an event. Suppose a total of 410 people exit the event. What is the probability that the 57th person to exit the event is selected for the poll?

   (A) $\frac{1}{317}$.

   (B) $\frac{1}{20}$.

   (C) $\frac{1}{410}$

   (D) Cannot be determined with the information provided.

   Answer is (B). The 57th person will be selected if and only if number 17 is chosen as the starting point, out of the integers 1 to 20. But since the selection of the starting point is random (meaning every number has the same chance of being selected), each integer from 1 to 20 has a probability of $\frac{1}{20}$ to be selected, so the probability of the 57th person being selected is also $\frac{1}{20}$.

10. Adrian takes an instant test for a viral disease. On the box of the test kit, it is mentioned that both the sensitivity and specificity of the test is 0.99. Upon checking the Ministry of Health website, he also finds that the disease affects 0.1% of Singapore residents. What is the probability (rounded to 2 decimal places) that Adrian has the disease if he obtains a positive test result from the test kit?

    (A) 0.01.

    (B) 0.99.

    (C) 0.09.

    (D) 0.91.

    Answer is (C). We can set up a contingency table for this scenario as follows:

	Test		
	Positive	Negative	Row total
Disease	99	1	100
No disease	999	98901	99900
Column total	1098	98902	100000

    From the table, we see that the conditional probability P(Disease | Test positive) $= \frac{99}{1098} = 0.09$.

11. There are 3 families X, Y and Z. The families have 2 children each. Family X has 1 boy and 1 girl. Family Y has only 2 girls and Family Z has only 2 boys. 1 child is randomly selected among the 6 children. If the selected child is a boy, what is the probability that he is from family X?

    (A) 0.

    (B) $\frac{1}{6}$.

    (C) $\frac{1}{3}$.

    (D) $\frac{1}{2}$.

    Anser is (C). The required probability is the conditional probability

$$P(\text{child from family X} \mid \text{child is a boy}) = \frac{P(\text{child is a boy from family X})}{P(\text{child is a boy})} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.$$

12. We wish to deploy a certain number of sensors around a particular area so as to detect intruders moving through the area. We may assume that the sensors function independently and each has probability 0.9 of detecting an intruder in the area. We would like to achieve at least 99.5% success rate of detecting an intruder using the sensors. What is the minimum number of sensors we need to deploy in order to achieve this target?

    (A) Two.

    (B) Three.

    (C) Four.

    (D) Target cannot be achieved.

    Answer is (B). To achieve at least 99.5% success rate of detecting an intruder using the sensors, it means that the probability of **not** detecting an intruder must be less than 0.5% (= 0.005). Any individual sensor has a probability of 0.1 of not detecting an intruder. Since sensors function independently, two sensors would fail to detect an intruder with probability $0.1 \times 0.1 = 0.01$, which does not meet the target. Three sensors would fail to detect an intruder with probability $0.1 \times 0.1 \times 0.1 = 0.001$ and this meets the target.

13. Suppose A and B are two events. Which of the following statements is/are true?

    (I) $P(A \text{ and } B)$ is always less than or equal to $P(A)$.

    (II) $P(A \mid B)$ is always less than or equal to $P(A)$.

    (A) Only (I).

    (B) Only (II).

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

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

    Answer is (A). Since the event "A and B" requires event A to also occur, $P(A \text{ and } B)$ is always less than or equal to $P(A)$. For statement (II), let us consider A as the event of drawing a black card from a typical deck of 52 playing cards and B as the event of drawing a Spade. Note that $P(A \mid B) = 1$ since Spade is a black card while $P(A) = \frac{26}{52} = \frac{1}{2}$. So statement (II) is not true in this case.

14. Let A and B be events of a sample space. Consider the following statements.

    (I) $P(A) + P(\text{not } A) = 1$.

    (II) $P(A) = P(A|B) + P(A \mid \text{ not } B)$.

    Which of the statements must be true?

    (A) Only (I).

    (B) Only (II).

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

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

    Answer is (A). Events A and (not A) are mutually exclusive since they cannot happen together. Furthermore, any outcome always belongs to either A or (not A), therefore the total probabilities will add up to 1. Hence, statement (I) is true. Statement (II) is not true. What is true is that $P(A)$ will always be between $P(A \mid B)$ and $P(A \mid \text{ not } B)$. This is an analogue of the basic rule of rates for probability.

    If you are still not convinced, consider the following example. Consider a poker deck of 52 cards. One card is drawn randomly from the deck. Let A denote the event of drawing an ace and B denote the event of drawing a number from Ace to 10.

    $P(A) = \frac{1}{13}$, but $P(A \mid B) + P(A \mid \text{ not } B) = \frac{1}{10}$, therefore the two probabilities are not equal.

15. A test comprises three multiple choice questions. The first question provides two options, and the remaining two questions provide 3 options each. Each question has a unique answer. Suppose that a student (who did not study at all for the test) picks an option randomly for each question, and that these picks are independent of each other. Which of the following is closest to the probability that this student gets exactly 1 out of 3 questions correct?

    (A) 0.22.

    (B) 0.33.

    (C) 0.44.

    (D) 0.55.

    Answer is (C). The event consists of 3 mutually exclusive parts, namely

    (i) First question correct, second and third questions wrong.

    (ii) Second question correct, first and third questions wrong.

(iii) Third question correct, first and second questions wrong.

The probability of (i) is $(\frac{1}{2})(\frac{2}{3})^2$. The probability of (ii) is $(\frac{1}{3})(\frac{1}{2})(\frac{2}{3})$. The probability of (iii) is the same as that in (ii). Since the three parts are mutually exclusive, the probability that the student gets exactly 1 out of 3 questions correct is

$$(\frac{1}{2})(\frac{2}{3})^2 + (\frac{1}{3})(\frac{1}{2})(\frac{2}{3}) + (\frac{1}{3})(\frac{1}{2})(\frac{2}{3}) \approx 0.44.$$

16. A bag contains four balls numbered 1, 2, 3 and 4. In a game, a ball is drawn once at random from the bag to have its number read. Next, a fair coin is tossed that number of times independently. The discrete random variable X is the number of heads observed from the coin tosses.

    Fill in the blank in the following statement:

    The probability of X being 0 is _____ (give your answer correct to 2 decimal places).

    Answer is 0.23. $P(X = 0)$ occurs when we have 0 heads recorded, regardless of the number on the ball drawn. We have 4 scenarios, when the ball drawn is numbered 1, 2, 3 and 4, respectively.

$$P(X = 0 \text{ and ball drawn is numbered 1}) = \frac{1}{4} \times \frac{1}{2}.$$

$$P(X = 0 \text{ and ball drawn is numbered 2}) = \frac{1}{4} \times \frac{1}{4}.$$

$$P(X = 0 \text{ and ball drawn is numbered 3}) = \frac{1}{4} \times \frac{1}{8}.$$

$$P(X = 0 \text{ and ball drawn is numbered 4}) = \frac{1}{4} \times \frac{1}{16}.$$

    Summing up the probabilities of all the scenarios, we yield $P(X = 0) = \frac{15}{64}$ which is 0.23 correct to 2 decimal places.

17. For two events A and B, $P(A) = 0.5$ and $P(B) = 0.2$. Which of the following statements is/are always true? Select all that apply.

    (A) $P(A \text{ or } B) = 0.7$ if A and B are independent events.

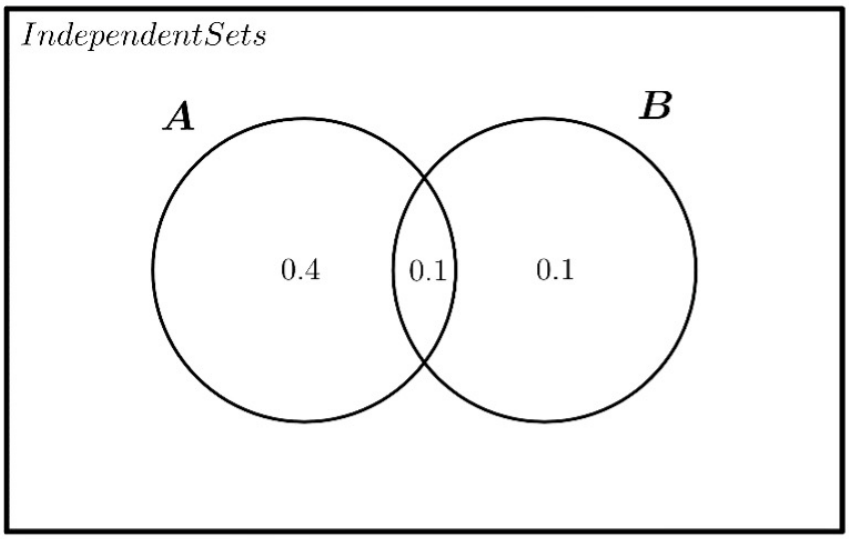
    (B) $P(A \text{ or } B) = 0.7$ if A and B are mutually exclusive events.

    (C) $P(A \text{ and } B) = 0.1$ if A and B are independent events.

    (D) $P(A \text{ and } B) = 0.1$ if A and B are mutually exclusive events.

    (E) $P(A \text{ and } B) = 0$ if A and B are mutually exclusive events.

    (B), (C) and (E) are correct. For mutually exclusive events, $P(A \text{ and } B) = 0$ and $P(A \text{ or } B) = P(A) + P(B) = 0.7$. For independent events, $P(A \text{ and } B) = P(A) \times P(B) = 0.1$ and $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.6$ (or derived via the intuitive Venn diagram shown below).



18. A bag consists of 10 balls. 4 of the balls are yellow while the remaining are green. 2 balls are drawn at random from the bag, one at a time with every ball having the same chance of being chosen on each draw. Let A be the event that the first ball drawn is yellow. Let B be the event that the second ball drawn is green. Which of the following statements is true?

    (A) If the balls are drawn **without replacement**, events A and B are independent and mutually exclusive.

    (B) If the balls are drawn **without replacement**, events A and B are independent but not mutually exclusive.

    (C) If the balls are drawn **with replacement**, events A and B are independent and mutually exclusive.

    (D) If the balls are drawn **with replacement**, events A and B are independent but not mutually exclusive.

    Answer is (D). A and B are not mutually exclusive regardless of whether the draws are made with replacement or not. If the balls are drawn **without replacement**, then $P(B \mid A) = \frac{2}{3}$ because after the first ball drawn is yellow, there are 9 balls left of which 6 are green. On the other hand, $P(B) = \frac{3}{5}$. To see why, observe that B can happen in 2 ways. The first way is when the first ball drawn is yellow and the second is green. The other way is when both balls drawn are green. Therefore the events are not independent. However, if the balls are drawn with replacement then $P(B \mid A) = \frac{3}{5} = P(B)$ so the events are independent since after each draw, the ball is replaced so we always have 10 balls of which 6 are green.

19. Based on a random sample of 200 staff members in NUS, the 95% confidence interval for the proportion of all NUS staff who went on vacation for at least 5 days in 2018 is $(0.33, 0.59)$. Which of the following statements must be true?

    (I) If another sample of size 500 is drawn using the same sampling method, for the same confidence level, the confidence interval will be wider than $(0.33, 0.59)$.

    (II) A maximum of 59% of all NUS staff went on vacation for at least 5 days in 2018.

    (A) Only (I).

    (B) Only (II).

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

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

Answer is (C). Statement (I) is false since a bigger sample size will likely result in a narrower confidence interval for the same confidence level. Statement (II) need not be true as the interval $(0.33, 0.59)$ is not guaranteed to contain the population proportion.

20. A researcher takes a random sample from Country X's population to estimate its unemployment rate. From the sample, the researcher obtains the 95% confidence interval for the population unemployment rate to be between 0.18 and 0.22.

    Which of the following statements correctly interprets the results?

    (A) If many samples of the same size were collected using the same procedure, and their respective confidence intervals calculated in the same way, about 95% of these samples will have the sample unemployment rate lie between 0.18 and 0.22.

    (B) If many samples of the same size were collected using the same procedure, and their respective confidence intervals calculated in the same way, about 95% of these samples will have the population unemployment rate lie between 0.18 and 0.22.

    (C) If many samples of the same size were collected using the same procedure, and their respective confidence intervals calculated in the same way, about 95% of these samples will have the sample unemployment rate lie within the samples' respective confidence intervals.

    (D) If many samples of the same size were collected using the same procedure, and their respective confidence intervals calculated in the same way, about 95% of these samples will have the population unemployment rate lie within the samples' respective confidence intervals.

    Answer is (D). By the interpretation of confidence intervals via repeated sampling, we conclude that about 95% of the samples will contain the population parameter (in this case the population unemployment rate) within their respective confidence intervals.

21. A 95% confidence interval, constructed from a random sample, for the population mean number of children per household in Country Z is $(1.21, 4.67)$. Which of the following statements is/are true? Select all that apply.

    (A) The probability that the population mean number of children per household in Country Z lies between 1.21 and 4.67 is 0.95.

    (B) We are 95% confident that the population mean number of children per household in Country Z lies between 1.21 and 4.67.

    (C) 95% of all samples of the same size and sampling procedure should have sample mean number of children per household between 1.21 and 4.67.

    (D) 95% of all households in Country Z have between 1.21 and 4.67 children.

    (E) If we take 100 different samples of the same size using the same sampling procedure and compute the confidence interval for each sample in the same way, approximately 95 of the intervals will contain the true population mean.

    (B) and (E) are correct. It is wrong to say that the probability of the population mean lies between 1.21 and 4.67 is 0.95. The population mean is **either** between 1.21 and 4.67 **or it is not**. There is no probability/chance involved. Confidence intervals are constructed based on sample means and when 100 intervals are constructed from the 100 samples, all the 100 intervals should capture their own sample means. Lastly, confidence intervals give us information about the population mean and not about individual households.

22. Two different random samples (call them Sample 1 and Sample 2) of size 100 each were chosen from a population of 10000 people. For Sample 1, the 95% confidence interval (call this Interval 1) for the population mean height was calculated. For Sample 2, the 99% confidence interval (call this Interval 2) for the population mean height was calculated. Which of the following statements is/are correct? Select all that apply.

    (A) If the population mean height lies in Interval 1, then it must lie in Interval 2.

    (B) If the population mean height lies in Interval 2, then it must lie in Interval 1.

    (C) The population mean height must lie in at least one of the two confidence intervals.

    (D) It is possible that the population mean height does not lie in both Intervals 1 and 2.

    (E) It is possible that the population mean height lies in both Intervals 1 and 2.

    (D) and (E) are correct. Any confidence interval constructed from any sample, regardless of the siginifcance level, may or may not contain the population parameter.

23. A random sample of size 500 is taken from a population of 10000 people of age 50. From the sample, a 95% confidence interval for the population mean weight is constructed. Which of the following statements is/are correct?

    (I) The confidence interval will always contain the sample mean weight.

    (II) If many samples of the same size are collected using the same sampling method, about 5% of the confidence intervals from these samples will not contain the population mean weight.

    (A) Only (I).

    (B) Only (II).

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

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

    Answer is (C). A confidence interval contructed from a sample will always contain the sample mean (while we cannot be sure if it will contain the population mean). For statement (II), the interpretation of confidence intervals via repeated sampling tells us that about 95% of the confidence intervals from these samples will contain the population mean, so about 5% of the intervals will not contain the population mean.

24. A 99% confidence interval for the mean height (in meters) of NUS students is $[1.58, 1.80]$. It is constructed using a random sample of 100 students. Using the same sample, which of the following is a plausible 95% confidence interval for the mean height?

    (A) $[1.64, 1.86]$.

    (B) $[1.61, 1.77]$.

    (C) $[1.48, 1.89]$.

    (D) $[1.63, 1.85]$.

    Answer is (B). From the 99% confidence interval [1.58, 1.80], we can infer that the sample mean is $\frac{(1.80+1.58)}{2} = 1.69$ and the margin of error is $\frac{(1.80-1.58)}{2} = 0.11$.

    As we use the same sample to calculate the 95% confidence interval, the sample mean remains the same but the margin of error decreases as the confidence level is lower. Therefore, the lower bound of the 95% confidence interval should be larger than the lower bound of the 99% confidence interval. Similarly, the upper bound of the 95% confidence interval should be smaller than the upper bound of the 99% confidence interval.

    By checking the options given, [1.61, 1.77] is the only plausible answer.

25. A coin manufacturer claims that he has produced a biased coin with $P(H) = 0.4$ and $P(T) = 0.6$, where $P(H)$ denotes the probability of the coin landing on heads and $P(T)$ denotes the probability of the coin landing on tails. Out of 10 independent tosses, Brad observes 8 heads and 2 tails. Based on these data, he decides to do a hypothesis test to see if there is enough evidence to reject the manufacturer's claim. Which of the following statements should he adopt as his null hypothesis?

    (A) $P(H) = 0.4$.

    (B) $P(H) = 0.5$.

    (C) $P(H) = 0.8$.

    (D) $P(H) = 0.6$.

    Answer is (A). Whenever we want to gather evidence to reject a claim, the null hypothesis should maintain that the claim is true. In this case, it should state $P(H) = 0.4$. It should not be based on the data, so $P(H) = 0.8$ is wrong. The other options are irrelevant.

26. Suppose we want to test if a coin is biased towards heads. We decide to toss the coin 10 times and record the number of heads. We shall assume the independence of coin tosses, so that the 10 tosses constitute a probability experiment.

    Let X denote the number of heads occurring in 10 tosses of the coin. We will carry out a hypothesis test with X as the test statistic. Let H be the event that the coin lands on head, in a single toss. We set our hypotheses to be

    - $H_0 : P(H) = 0.5$,
    - $H_1 : P(H) > 0.5$.

    Suppose in our execution of the 10 tosses, we observe 4 heads. This means $X = 4$ is the test result we observe.

    Recall the definition of p-value to be the probability of obtaining a test result at least as extreme as the one observed, assuming the null hypothesis is true. What is the range of test results "at least as extreme as the one observed", in this scenario?

    (A) $0 \leq X \leq 4$.

    (B) $4 \leq X \leq 10$.

    (C) $0 \leq X \leq 5$.

    (D) $5 \leq X \leq 10$.

    Answer is (B). Note that in the context of p-value computation, "at least as extreme" is interpreted as "at least as favourable to the alternative hypothesis". In this scenario, the greater the value X assumes, the more favourable the case is to the alternative hypothesis. Hence, the answer should be all values of X greater than or equal to the observed value.

27. A group of students wants to find out if there is any association between staying in a hall and being late for class in NUS in a particular month. If students are late for at least 5 classes, they are considered "late for class" in that month. After collecting a random sample of 1000 students, they found that 200 out of 350 students who stay in a hall are late for class, while 390 out of 650 students who do not stay in a hall are late for class.

    A chi-squared test was done to test for association between staying in a hall and being late for class at 5% level of significance. The p-value derived from the chi-squared test is 0.3809.

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

    (A) There is a positive association between staying in a hall and being late at the sample level.

    (B) There is a negative association between staying in a hall and being late at the sample level.

(C) Since the p-value is more than 0.05, we can conclude that there is an association between staying in a hall and being late at the population level.

(D) Since the p-value is more than 0.05, we cannot conclude that there is an association between staying in a hall and being late at the population level.

(B) and (D) are correct. To check for association between staying in hall and being late at the sample level, we compare

$$\begin{aligned} &\text{rate(Late for class} \mid \text{Staying in hall)} = \frac{200}{350} \\ < \quad &\frac{390}{650} = \text{rate(Late for class} \mid \text{Not staying in hall)} \end{aligned}$$

Hence, we see that staying in hall is negatively associated with being late at the sample level. As we are doing a chi-squared test to see if there is an association at the population level, recall that the null hypothesis states that there is no association at the population level, while the alternative hypothesis states that there is an association at the population level. Since the p-value is greater than the level of significance, we do not reject the null hypothesis. Hence, we cannot conclude that there is an association between staying in hall and being late at the population level.

28. 25 mothers were each allowed to smell two articles of infant's clothing. Each of them was then asked to pick the one which belongs to her infant. They were successful in doing so 72% of the time. You want to show that this has not happened by chance and mothers can indeed recognise the smell of their children. To test such a hypothesis, what should the null and alternative hypotheses be?

(A) H_0: P(Success)= 0.5; H_1: P(Success)> 0.5.

(B) H_0: P(Success)= 0.72; H_1: P(Success)> 0.72.

(C) H_0: P(Success)= 0.5; H_1: P(Success)< 0.5.

(D) H_0: P(Success)= 0.72; H_1: P(Success)< 0.72.

Answer is (A). To show something does not happen by chance, we should have the null hypothesis implying that it happens by chance, and the alternative hypothesis implying that it does not. This is so that we can gather evidence from a probability experiment to reject the null for the alternative. In this case, the null hypothesis is that the chance that a mother is able to recognise her child smell correctly is 50%. The alternative hypothesis should be in line with what we want to show, hence the alternative hypothesis is H_1 : P(Success)> 0.5.

29. A hypothesis test is done to find out whether vaccine X prevents cancer in a population of dogs, where cancer affects 10% of dogs. A random sample of 100 puppies was selected for the study. All 100 puppies received vaccine X and we observed them for their entire lifetimes. 5 of these puppies evertually had cancer. The null hypothesis is

$$H_0\text{: Vaccine X has no effect on cancer in the population.}$$

Then the p-value is

(A) the probability that vaccine X is effective.

(B) the probability that vaccine X is not effective.

(C) the probability that the hypothesis will be rejected.

(D) the probability that 5 puppies out of 100 have cancer, given that the probability of cancer is 0.1.

(E) the probability that 5 or less puppies out of 100 have cancer, given that the probability of cancer is 0.1.

Answer is (E). The p-value is the probability of obtaining a test result that is equal to or more extreme than the observed, given that the probability of cancer in dogs is 0.1. Here, an "equal" or more extreme result is 5 or less puppies out of 100 having cancer.

30. Mandy and Sue were playing a game - Sue draws a random card from a deck of five cards (red, blue, yellow, green and black) and hides it out of sight from Mandy. Mandy will try to guess the colour of that drawn card. Mandy wins if she can correctly guess the colour of the drawn card. Otherwise, Mandy loses.

    They played 5 rounds of the game, and Mandy won 4 out of 5 games. Sue is surprised that Mandy won so many times, and suspects Mandy may have some method to detect the colour of the cards instead of just guessing the colour. Based on the above, she wants to conduct a hypothesis test, with the null hypothesis:

    $$P(\text{Mandy guesses a card colour correctly}) = 0.2.$$

    What event(s) need to be considered in the calculation of the p-value? Select all that apply.

    (A) Event: Mandy losing 5 games out of 5.

    (B) Event: Mandy losing any 4 games out of 5.

    (C) Event: Mandy losing any 3 games out of 5.

    (D) Event: Mandy losing any 2 games out of 5.

    (E) Event: Mandy losing any 1 game out of 5.

    (F) Event: Mandy losing 0 games out of 5.

    (E) and (F) are correct. p-value is the probability of obtaining a test result at least as extreme as the result observed, assuming the null hypothesis is true. In this case, the observed result is Mandy losing 1 game out of 5. Then, an event giving rise to a more extreme result is Mandy losing 0 games.