1. For the survey question "How many universities did you apply to?", the data obtained is summarized below in a frequency table. A total of 47 students answered this question.

Number of universities	0	1	2	3	4	5	6	9
Frequency	6	6	19	8	2	4	1	1

(a) (2 points) Calculate the sample mean.

(b) (3 points) Find the five number summary and draw the box-plot (without outliers).

(c) (2 points) Are there any outliers in this data? Show you work.

(d) (1 point) Is the distribution symmetric, and if not, which way is it skewed?

2. (3 points) It was estimated that the average weight of the brain (among adults) is 1350g, with a standard deviation of 160g. Use Chebyshev's Theorem to construct an interval that is guaranteed to contain at least 90% of the adult brain weights.

3. Suppose A and B are *disjoint* events, and B and C are *independent* events. Also suppose you know the probabilities:

$$P(A) = 0.23, \quad P(A \cup B) = 0.61, \quad P(B \cap C) = 0.30, \quad P(A \cap C) = 0.15.$$

Calculate the following probabilities:

(a) (1 point) $P(B)$

(b) (2 points) $P(B \cup C)$

(c) (1 point) $P(A \cup B')$

(d) (2 points) $P(A' \cap B' \cap C)$

(e) (1 point) $P(C|A)$

4. (4 points) A study found that, among males, the chance of developing lung cancer is 1 out of 13. Among the males who get lung cancer, 90% of them are regular smokers. Among those who will never get lung cancer, 20% of them are smokers.

Given that I am not a smoker, what is the probability I will get lung cancer in my lifetime?

5. Suppose we attempt to predict the long-term weather randomly, for 7 consecutive days in the future. In a bag, we put a total of 20 pieces of paper: on 12 of them we have drawn a sun, and on the remaining 8 we have drawn a cloud with rain.

Then, we pick at random 7 pieces of paper from the bag: the first gives the weather prediction for the first day, the second is the weather prediction for the second day, and so on.

(a) (2 points) Assuming the draws are done *with replacement*, find the probability that we will predict sun on the first 2 days, and clouds during the remaining 5 days.

(b) (2 points) Assuming the draws are done *without replacement*, find the probability that we will predict 4 consecutive sunny days (not necessarily starting at the beginning of the sequence), and cloudy weather for the remaining 3 days.

6. An insurance company offers its policyholders a number of different payment options. Each customer can choose to pay every month, every 3 months, every 4 months, every 6 months or every year. For a randomly selected customer, let X denote the number of months for the selected payment method. Below is the **cumulative** distribution function $F(x)$.

$$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 0.3 & \text{if } 1 \le x < 3 \\ 0.4 & \text{if } 3 \le x < 4 \\ 0.45 & \text{if } 4 \le x < 6 \\ 0.6 & \text{if } 6 \le x < 12 \\ 1 & \text{if } x \ge 12 \end{cases}$$

(a) (2 points) Find the probability distribution $p(x)$.

(b) (3 points) Calculate the expected value and the variance of X.

(c) (2 points) Find the probability that X is within one standard deviation of the expected value.

7. (3 points) Consider the following continuous probability density function.

$$f(x) = \begin{cases} x - 1 & \text{for } 1 \leq x < 2 \\ 2 - \dfrac{x}{2} & \text{for } 2 \leq x < A \\ 0 & \text{elsewhere.} \end{cases}$$

Find the value of the parameter A that would make $f(x)$ a valid probability density function.
(*Hint: there is only one answer.*)

8. When inspecting a section of the Champlain bridge, there are two kinds of cracks: hairline cracks, and major cracks. Suppose the number of hairline cracks follows a Poisson distribution with an average of 4 cracks per 30-foot section, and the number of major cracks follows a Poisson distribution with an average of 0.5 cracks per 30-foot section.

(a) (2 points) If the inspector looks at a 100-foot section of the bridge, what is the probability that he will find exactly 10 hairline cracks?

(b) (3 points) For a particular 30-foot section, what is the probability that it contains a total of 2 cracks of any kind? Assume that the number of hairline cracks and the number of major cracks are independent.

9. According to the data obtained by the National Safety Council in the United States, 26% of all car crashes in 2014 were caused by using a cellphone while driving.

(a) (2 points) What is the probability that, for the next 20 car crashes, exactly 5 of them will be caused by using a cellphone?

(b) (4 points) What is the probability that, for the next 1000 car crashes in the United States, at least 250 and at most 300 of them will be caused by using a cellphone? Use a normal approximation, and verify first that it is valid to use such an approximation.

10. Suppose that the lifespan of a car battery follows a normal distribution with an average of 5 years, and a standard deviation of 1.2 years.

(a) (2 points) Find the probability that a given car battery will last more than 6 years.

(b) (2 points) If we take a random sample of 10 car batteries, what is the probability that the average lifespan of these batteries will exceed 6 years?

(c) (2 points) Again using a sample of 10 batteries, what is the probability that exactly 5 of them will last more than 6 years?

11. (3 points) In a fishing event, a small lake is populated with 75 trouts, among which 25 are tagged. Each participant is allowed to capture 5 fish during the day (the fish are **not** put back into the lake). For each tagged fish, the participant wins 10$, and for each non-tagged fish, the participant wins 2$. However, it costs 25$ to participate in this event. Find the expected value and variance of a participant's net earnings. (Assume for simplicity that there's only one participant, so nobody else is fishing at the same time.)

12. Someone wishes to estimate the average wrist circumference (in millimetres) of CEGEP students. From a sample of 44 measurements, a 98% confidence interval was constructed with the following result: [154.49mm, 164.71mm].

(a) (2 points) Without making any calculations, would a 92% confidence interval be wider or narrower? Explain.

(b) (3 points) Construct a 92% confidence interval for the average wrist circumference of students.

13. (4 points) Using a sample of 6 left-handed students in this class, the average R-score was found to be 28.59, with a standard deviation of 4.13. Assume that the R-score is normally distributed. Construct a 95% confidence interval for the true standard deviation of the R-score among left-handed students, and interpret your answer.

14. (3 points) Find the 5^{th} percentile for the following distributions:

(a) $N(-5, 4)$

(b) t_{24}

(c) $F_{5,10}$

15. (4 points) The IQ score is claimed to have an average of 100, with a standard deviation of 15. In order the disprove this claim, a researcher is planning to test the null hypothesis $H_0 : \mu = 100$ against the alternative $H_a : \mu \neq 100$, with a 5% significance level and a random sample of 81 individuals.

Find $\beta(102)$, the probability of committing a type II error if the true average IQ score is 102.

16. (5 points) At École Polytechnique (an engineering school), the proportion of female students is evolving with time. In the Fall semester of 2013, there were 141 female students out of 897 registered in the first year of the mechanical engineering program. Five years before, in the Fall of 2008, there were 105 female students out of 813 in mechanical engineering.

Test the claim that the proportion of female students in mechanical engineering has increased in the last five years. Use a 10% significance level, and follow the rejection region approach.

17. Sintering is the process of forming a solid mass of material starting from powder, without melting it. For sintering copper, two procedures are tested on six different samples of powder. The measurement of interest is the porosity of the resulting material. The results are shown below.

Powder Type	1	2	3	4	5	6
Porosity - Procedure I	21	27	18	22	26	19
Porosity - Procedure II	23	26	21	24	25	16

(a) (5 points) Perform a test for the claim that there is a significant difference in the porosity between the two procedures. Use a 5% significance level and follow the P-value approach.

(b) (1 point) Interpret your conclusion from part (a).

18. A small company has 15 female employees and 10 male employees. Every week, the company holds a lottery in which two different employees are chosen at random; the winners are allowed to leave one hour early on Friday afternoon. For a given week, let X denote the number of female employees winning the lottery.

(a) (3 points) Find the probability distribution of X assuming the lottery is truly random.

(b) (5 points) Suppose that, after 52 weeks of running this lottery, the number of females employees who won per week was tabulated. The observed frequencies are shown below.

Number of female winners	0	1	2	Total
Number of weeks	4	28	20	52

Would this data indicate that the lottery is not done randomly? Perform a test with a 5% significance level and following the rejection region approach.
(*Hint: you will need to use your answer from part (a).*)

(c) (1 point) Was it valid to apply the test you used in part (b) in this situation?

19. In a paper by Giovanni Tanda (2011) titled "*Prediction of marathon performance time on the basis of training indices*", the author analyzed the relationship between the time to complete a marathon race and some parameters of the training program prior to the race.

Let X denote the average workout distance during eight weeks prior to the race (in km/week) and Y denote the finishing time at the marathon (in minutes). Data for a sample of 22 Italian runners was gathered, and the following summary was obtained.

$$\sum x_i = 1450, \qquad \sum y_i = 4202, \qquad \sum x_i^2 = 100851, \qquad \sum y_i^2 = 805606, \qquad \sum x_i y_i = 273345.$$

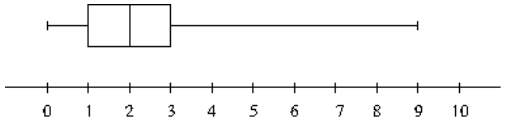
(a) (3 points) Calculate the correlation coefficient and interpret.

(b) (3 points) Find the equation of the least-squares regression line, and also estimate the variance of the error term.

(c) (2 points) Suppose a beginner is training for his first marathon. He has been running an average of 50 km/week prior to the race. What is the probability that he will complete the marathon with a time below 3 hours and a half? (3.5 hours = 210 minutes)

ANSWERS:

1. (a) 2.3617
   (b) five-number summary: 0, 1, 2, 3, 9



   (c) 9 is an outlier
   (d) not symmetric, positively skewed
2. $[844.0356, 1855.9644]$
3. (a) 0.38 (b) 0.8695 (c) 0.62 (d) 0.3395 (e) 0.6522
4. 0.01031
5. (a) 0.003686 (b) 0.04087
6. (a) $p(1) = 0.3, \quad p(3) = 0.1, \quad p(4) = 0.05, \quad p(6) = 0.15, \quad p(12) = 0.4$
   (b) $E(X) = 6.5, \quad V(X) = 22.75$
   (c) 0.3
7. $A = 2.5858$
8. (a) 0.07926 (b) 0.1124
9. (a) 0.2013 (b) 0.7746
10. (a) 0.2033 (b) 0.0041 (c) 0.02809
11. $E(Y) = -1.6667, \quad V(Y) = 67.2673$
12. (a) narrower, smaller chance of capturing the true average
    (b) $[155.762, 163.438]$
13. $[2.5780, 10.1306]$, there is a 95% chance that the true standard deviation in the R-score among all left-handed students is between 2.5780 and 10.1306.
14. (a) -8.29 (b) -1.711 (c) 0.21097
15. 0.7756
16. test on difference of proportions (upper-tail), $z^* = 1.6474$, reject H_0
17. (a) test on paired difference (two-tail), $t^* = -0.3492$, P-value > 0.20, fail to reject H_0
    (b) there is insufficient evidence to claim there is a significant difference between the porosity produced by the two methods, at the 5% significance level.
18. (a) $p(0) = 0.15, \quad p(1) = 0.5, \quad p(2) = 0.35$
    (b) goodness-of-fit test (upper-tail), $\chi^{2,*} = 2.1831$, fail to reject H_0
    (c) yes, since the expected frequencies f_e are all greater than 5
19. (a) $r = -0.9019$, very strong negative linear relationship
    (b) $Y = 235.9764 - 0.6824X, \quad s = 5.3094$
    (c) 0.9370