SM-1402/CU-0304 Basic Statistics – Semester 2 2019/20

This assignment counts for 20% of your summative assessment (10% for CU-0304 students). Attempt all questions and submit your solutions to my pigeonhole by 12:00 pm on **Tue, 3 Mar 2020**. The total marks attainable is 50. Late submissions will be penalised.

1. The following two sets of data represent the lengths (in minutes) of students attention spans during a one-hour lecture.

Statistics class

01	43	16	28	27	25	26	25	22	26
47	40	14	36	23	32	15	31	19	25
21	07	28	49	31	22	24	26	14	45
38	48	36	22	29	12	32	11	34	42
55	27	06	23	42	21	58	23	35	13

Economics class

60	39	30	41	37	27	38	04	25	43
58	60	21	53	26	47	08	51	19	31
29	21	31	60	48	30	28	37	07	60
50	60	51	24	41	03	37	14	46	60
60	48	25	32	59	11	60	28	54	18
60	42	04	26	60	41	60	11	43	28

(a) (7 marks) Construct histograms for each of these datasets and use these to comment on comparisons between the two distributions.

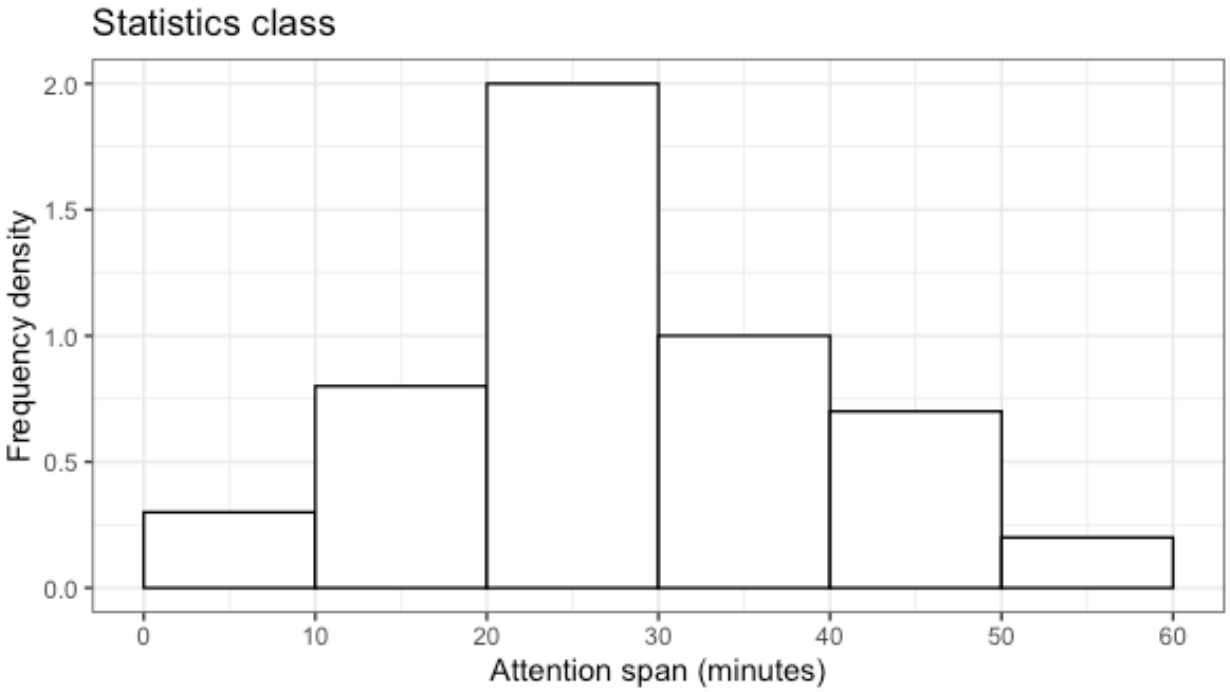
Solution: It seems natural to use interval width of 10 minutes. For the Statistics class, we have:

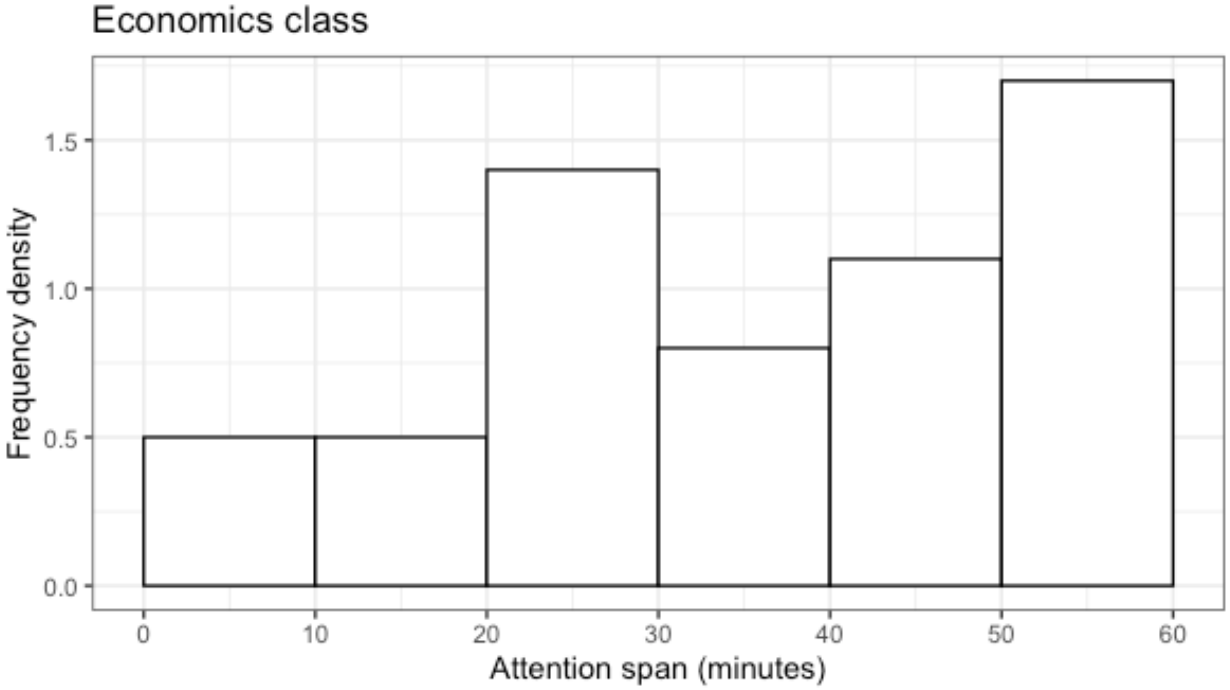
Attention span (minutes)	Frequency	Interval width	Frequency density
$0 \leq x < 10$	3	10	0.3
$10 \leq x < 20$	8	10	0.8
$20 \leq x < 30$	20	10	2
$30 \leq x < 40$	10	10	1
$40 \leq x < 50$	7	10	0.7
$50 \leq x < 60$	2	10	0.2

And for the Economics class,

Attention span (minutes)	Frequency	Interval width	Frequency density
$0 \leq x < 10$	5	10	0.5
$10 \leq x < 20$	5	10	0.5
$20 \leq x < 30$	14	10	1.4
$30 \leq x < 40$	8	10	0.8
$40 \leq x < 50$	11	10	1.1
$50 \leq x < 60$	17	10	1.7

Comment: Using "frequency" for the y-axis of the histogram would not be terribly wrong in this case, since all class widths are equivalent. Just make sure to label the axes appropriately.





(b) (3 marks) Compare the means and standard deviations of the two groups and comment. You may wish to use the following summary statistics:

	Statistics Class	Economics Class
Sum	1395	2225
Sum of squares	46713	100387

Solution: For the Statistics class, we have that $\bar{x} = \frac{1}{50}\sum x = 1395/50 = 27.9$ and $s^2 = \frac{1}{50}\sum x - \bar{x}^2 = 46713/50 - 27.9^2 = 155.85$. The standard deviation is $s = \sqrt{155.85} = 12.5$.

For the Economics class, $\bar{x} = \frac{1}{60}\sum x = 2225/60 = 37.1$ and $s^2 = \frac{1}{60}\sum x - \bar{x}^2 = 100387/60 - 37.1^2 = 296.706$. The standard deviation is $s = \sqrt{296.706} = 17.2$.

It seems that students in the statistics class have shorter attention span compared to their economics counterpart. However, the spread of attention span is much larger in economics class. This indicates that although the attention span in statistics is lower, students are more focused as a whole group, compared to economics.

2. An airline has a baggage weight allowance of 20 kg per passenger on its flights. It collected the following data on the weights of its passengers baggage. The data were taken from a sample of 100 passengers on a Brunei-Singapore flight. There were about 300 passengers on the flight and the 100 passengers were chosen at random.

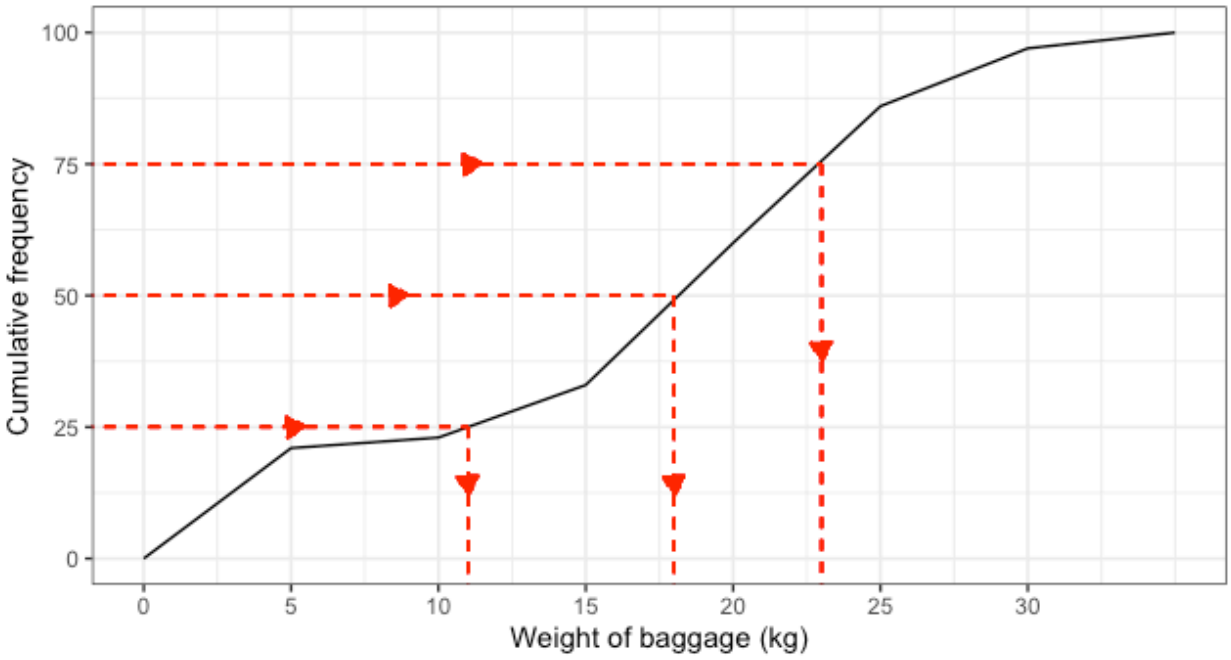
Weight of baggage (kg)	No. of passengers	Cum. Freq.
$x < 0$	0	0
$0 \leq x < 5$	21	21
$5 \leq x < 10$	2	23
$10 \leq x < 15$	10	33
$15 \leq x < 20$	27	60
$20 \leq x < 25$	26	86
$25 \leq x < 30$	11	97
$30 \leq x < 35$	3	100
$35 \leq x$	0	100

(a) (3 marks) Using a graph paper, draw a cumulative frequency polygon for the grouped data above. Show your working, i.e. the cumulative frequency table.

Solution: The table of interest is as above.

(b) (4 marks) Determine the median, lower quartile, upper quartile and IQR.

Solution: Draw lines at cumulative frequency equal to 25 (Q1), 50 (Q2) and 75 (Q3) to determine these values to be 11kg, 18kg and 23kg respectively. The inter-quartile range (IQR) is $23 - 11 = 12$ kg.



(c) (2 marks) Calculate the mean of the data.

Solution: The mean is obtained by multiplying the *midpoint* of the intervals by the frequency, summing them up, and then dividing by the total frequency.

$$\bar{x} = \frac{2.5 \times 21 + 7.5 \times 2 + 12.5 \times 10 + \cdots + 27.5 \times 11 + 32.5 \times 3}{100}$$
$$= 16.5 \text{ kg}$$

(d) (1 mark) The company uses the data to claim that "40% of airline passengers travel with baggage over the weight allowance". Explain whether or not you think this claim is valid. *Hint: Think about how the data were collected.*

Solution: From the cumulative frequency graph, the data show that 40% of passenger would in fact travel with less than 20 kg of baggage (weight allowance). This claim is wrong for this particular data set, i.e. passengers travelling on the Brunei–Singapore flights. It is also wrong to then infer this for *all* passengers since data collected is not representative of passengers flying to/from the other destinations within the airlines' routes.

3. In a sample of $n = 6$ objects, the mean of the data is 15 and the median is 11. Another observation is then added to the sample mean and this takes the value $x_7 = 12$.

(a) (2 marks) Calculate the new mean of the seven observations.

Solution: Indicate by $\bar{x}_n$ the mean of n observations. Thus, we have that $\bar{x}_6 = \frac{1}{6}\sum_{i=1}^{6} x_i = 15$, therefore $\sum_{i=1}^{6} x_i = 6 \times 15 = 90$. We can use this to calculate the mean of 7 observations, since

$$\begin{aligned}\bar{x}_7 = \frac{1}{7}\sum_{i=1}^{7} x_i &= \frac{1}{7}(x_1 + x_2 + \cdots + x_6 + x_7) \\ &= \frac{1}{7}\left(\sum_{i=1}^{6} x_i + x_7\right) \\ &= \frac{1}{7}(90 + 12) \\ &= 14.6\end{aligned}$$

(b) (3 marks) What can you conclude about the median of the seven observations?

Solution: Denote the order statistics as $x_{(k)}$, i.e. $x_{(k)}$ is the k'th largest value among all the observations. As there are 6 observations, the median is the average of the values of $x_{(3)}$ and $x_{(4)}$ by definition. There are several cases:

- Case 1: $x_{(3)} = x_{(4)} = 11$. In this case, adding a new observation $x_7 = 12$ will put it on the right of $x_{(3)}$ and $x_{(4)}$, so therefore the median will not change because the new median is now $x_{(4)} = 11$.
- Case 2: $10 \leq x_{(3)} < 11$ and $11 < x_{(4)} \leq 12$. In this case, adding a new observation $x_7 = 12$ will put also it on the right of $x_{(4)}$, and now making the new median $x_{(4)}$, which can now take any value in $(11, 12]$.
- Case 3: $x_{(3)} < 10$ and $x_{(4)} > 12$. In this case, adding a new observation $x_7 = 12$ will put it in between $x_{(3)}$ and $x_{(4)}$, therefore making the new median 12.

Therefore, the new median is any value between 11 and 12 inclusive.

4. (3 marks) The successful operation of three separate switches is needed to control a machine. If the probability of failure of each switch is 0.1 and the failure of any switch is independent of any other switch, what is the probability that the machine will break down?

Solution: You could use probability trees and work out the cases where the machine will break down, but a simpler way would be to compute the probability that the machine *will not* break down. There is exactly one instance where this occurs, and that is when all switches work. Since the switches are independent of each other, we have

$$P(\text{machine not break down}) = 0.9 \times 0.9 \times 0.9 = 0.729,$$

and therefore $P(\text{machine break down}) = 1 - P(\text{machine not break down}) = 1 - 0.729 = 0.271$.

5. (3 marks) In a certain states lottery, 48 balls numbered 1 through 48 are placed in a machine and six of them are drawn at random. If five of the six numbers drawn match the numbers that a player has chosen, the player wins a second prize of $1,000. Compute the probability that you win the second prize if you purchase a single lottery ticket.

Solution: The number of possible outcomes of the lottery drawing is $\binom{48}{6}$. In order to win the second prize, five of the six numbers on the ticket must match five of the six winning numbers; in other words, we must have chosen five of the six winning numbers and one of the 42 losing numbers. The number of ways to choose 5 out of the 6 winning numbers is given by $\binom{6}{5} = 6$, and the number of ways to choose 1 out of the 42 losing numbers is given by $\binom{42}{1} = 42$. Thus the number of favourable outcomes is then given by $6 \times 42 = 252$. So the probability of winning the second prize is $252/\binom{48}{6} = 0.0000205$.

If this was confusing, try think of a simpler example. Let's consider the case where there are 4 balls numbered 1 to 4, and the lottery draws 3 of these balls. How many possible combinations are there? Ans: $\binom{4}{3} = 4$. We can even list them out: 123, 124, 134, and 234. For arguments sake, let's say that the winning number is 123. Out of these three winning numbers, how many ways of choosing only 2 of the winning numbers? That would be 12x, 23x, and 13x, or in other words, $\binom{3}{2} = 3$. Now as for the 'x', it has to be *not* a winning number, meaning we have to choose it from the balls that were *never* picked. Since there are 4 balls altogether, 3 were picked, so $4-3 = 1$ remains, and in fact there is only one way of picking this number. So the total number of ways to obtain 2 out of 3 winning numbers in this case is $3 \times 1 = 3$.

6. (4 marks) Two fair dice are thrown. Determine the probability distribution of the absolute difference of the two dice, Y, and find its mean and variance. *Hint: If Z_1 is the score of dice 1 and Z_2 is the score of dice 2, $Y = |Z_1 - Z_2|$.*

Solution: The possible values that X can take are $0, 1, 2, 3, 4, 5$ only. The probability distribution table is

x	0	1	2	3	4	5
$P(X = x)$	6/36	10/36	8/36	6/36	4/36	2/36

The mean is given by

$$E(X) = 0 + 1(10/36) + 2(8/36) + 3(6/36) + 4(4/36) + 5(2/36) = 70/36 = 1.94.$$

To calculate the variance, calculate $E(X^2)$ first:

$$E(X^2) = 0 + 1^2(10/36) + 2^2(8/36) + 3^2(6/36) + 4^2(4/36) + 5^2(2/36) = 70/36 = 210/36.$$

So the variance is $\mathrm{Var}(X) = E(X^2) - E^2(X) = 210/36 - (70/36)^2 = 2.05$.

7. An examination consists of four multiple choice questions, each with a choice of three answers. Let X be the number of questions answered correctly when a student resorts to pure guesswork for each answer.

   (a) (1 mark) State which special probability distribution, together with its parameters, you would use to model X.

   Solution: $X \sim \text{Bin}(4, 1/3)$.

   (b) (4 marks) Tabulate the probability distribution function for X.

   Solution: A student can answer 0, 1, 2, 3, or 4 correct. Thus,

x	0	1	2	3	4
$P(X = x)$	0.198	0.395	0.296	0.099	0.012

   (c) (3 marks) If each question is worth one mark, what is the mean score and standard deviation of scores when a student completely guesses the answers to all four questions?

   Solution: $\text{E}(X) = np = 4 \times 1/3 = 1.33$, and $\text{Var}(X) = np(1-p) = 4 \times 1/3 \times 2/3 = 0.89$. Then $SD(X) = \sqrt{0.89} = 0.943$.

   (d) (2 marks) The examiner calculates a rescaled mark using the formula $Y = 10 + 22.5X$. Find the mean of Y.

   Solution: $\text{E}(Y) = \text{E}(10 + 22.5X) = 10 + 22.5\,\text{E}(X) = 10 + 22.5 \times 1.33 = 39.9$.

8. Over a period of time the number of break-ins per month in a given district has been observed to follow a Poisson distribution with mean 2.

   (a) For a given month, find the probability that the number of break-ins is

      i. (1 mark) Fewer than 2.

      Solution: Let X be the number of break-ins in a month. Then $X \sim \text{Poi}(2)$.

$$\begin{aligned} \therefore P(X < 2) &= P(X = 0) + P(X = 1) \\ &= e^{-2} + 2e^{-2} \\ &= e^{-2}(1 + 2) \\ &= 0.406 \end{aligned}$$

      ii. (1 mark) More than 4.

Solution:

$$\begin{aligned}
P(X > 4) &= 1 - P(X \leq 4) \\
&= 1 - \big\{P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)\big\} \\
&= 1 - e^{-2}\left\{1 + 2 + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!}\right\} \\
&= 1 - 7e^{-2} \\
&= 1 - 0.94734 = 0.0527
\end{aligned}$$

iii. (1 mark) At least 1, but no more than 3.

Solution: 'At least 1' means $\{X \geq 1\}$. 'No more than 3' means $\{X \leq 3\}$. So,

$$\begin{aligned}
P(1 \leq X \leq 3) &= P(X=1) + P(X=2) + P(X=3) \\
&= 2e^{-2} + \frac{2^2 e^{-2}}{2!} + \frac{2^3 e^{-2}}{3!} \\
&= e^{-2}(2 + 2^2/2! + 2^3/3!) \\
&= 5.333 \times e^{-2} \\
&= 0.722
\end{aligned}$$

(b) (2 marks) What is the probability that there will be fewer than ten break-ins in a six-month period?

Solution: Let Y be the number of break-ins in a six-month period. If there is, on average, 2 break-ins per month, therefore there is, on average, $2 \times 6 = 12$ break-ins in a six-month period. Thus, $Y \sim \text{Poi}(12)$.

$$\begin{aligned}
P(Y < 10) &= P(Y=0) + P(Y=1) + \cdots + P(Y=9) \\
&= e^{-12}\left\{1 + 12 + \frac{12^2}{2!} + \frac{12^3}{3!} + \frac{12^4}{4!} + \frac{12^5}{5!} + \frac{12^6}{6!} + \frac{12^7}{7!} + \frac{12^8}{8!} + \frac{12^9}{9!}\right\} \\
&= e^{-12}(13 + 72 + 288 + 864 + 2073.6 + 4147.2 + 7109.486 \\
&\qquad + 10664.229 + 14218.971) \\
&= 39450.49e^{-12} \\
&= 0.242
\end{aligned}$$

Comment: This was a very tedious calculation, but it is not impossible to do by hand. However, we can also use other means, such as looking up this probability from a computer, or statistical tables.