Homework 1

S1201, Summer 2022

100 points

Due July 14, 23:59PM EST

Notes: Please read the following instructions.

(i) Please submit a **single** PDF file on courseworks. You may scan your written solutions or directly write the solutions on your tablet.

(ii) Late submission will lead to some penalty: 20 points off for within 24 hours late, 40 points off for over 24 hours late, all points off for submission after the solutions are posted on courseworks.

(iii) You can discuss the problems with others. But a direct plagiarism will lead to zero point for this assignment, and this will be reported to the university.

(iv) Feel free to use calculators!

(v) There is one extra problem with 5 bonus points, which is optional. Therefore all points add up to 105, but the maximum score of this homework is 100.

(vi) You may need to use the trick of interpolations to calculate the quantiles in Problems 2 and 5. Remember to rank the data first in Problem 2 before you calculate anything. As I mentioned in class, I do not require everyone to master this trick very well. If you can do it, welcome to try by yourself first, then compare your results with the calculator [here]. Enter the data (no need to be ranked) and the percentage, click calculate, and follow the number given by "Definition 2". If you can't do it, just use the result from the calculator. No worries.

1. **$2 \times 6 = 12$ points** TRUE/FALSE questions. No explanations are needed.

   (a) The heights of the bars in a density histogram sum to 1. should be the area F.

   (b) When we draw histograms, the smaller bin widths we use, the better. F

   (c) Box plot can help identify the potential outliers. T

   (d) Sample median is more sensitive to extreme values than the sample mean. F

   (e) If three events A, B and C are mutually exclusive, then any two of them are mutually exclusive as well. T

   (f) Given an experiment and the sample space, the probability of an event depends on the samples we collect. F is deterministic !

2. **2+4+2+2=10 points** How many points do football teams score in the Super Bowl? Here are the total numbers of points scored by both teams in each of the first 20 Super Bowl games:

   44, 47, 23, 29, 30, 28, 21, 30, 22, 39, 46, 37, 67, 49, 37, 46, 43, 47, 55, 57

   (a) Find the median;

   (b) Find the quartiles (all four quartiles);

   (c) Write down the 5-number summary;

   (d) Calculate the standard deviation.

Solution: (a) $n = 20$ data points.

Rank the data:

21, 22, 23, 28, 29, 30, 30, 37, 37, 39, 43, 44, 46, 46, 47, 47, 49, 55, 57, 67

sample median $= \frac{1}{2}(39+43) = 41$

(b) ① 29 is $(100 \times \frac{4}{19})\% = 21.05\%$ quantile

30 is $(100 \times \frac{5}{19})\% = 26.32\%$ quantile

And $25\% = 25\% \times 21.05\% + 75\% \times 26.32\%$

By interpolation: 25% quantile $= 25\% \times 29 + 75\% \times 30 = 29.75$

Similarly: ②: 3rd quartile $= 47$

4th quartile $=$ maximum $= 67$

(c)

Min	Q1	Median	Q3	max
21	29.75	41	47	67

(d) $\bar{x} = 39.85$, $n = 20$

x_i	21	22	23	28	29	30	30	37	37	39	43	44	46	46	47	47	49	55	57	67
$x_i - \bar{x}$	-20	-19	-18	-13	-12	-11	-11	-4	-4	-2	2	3	5	5	6	6	8	14	16	26
$(x_i - \bar{x})^2$	400	361	324	169	144	121	121	16	16	4	4	9	25	25	36	36	64	196	256	676

SD $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{19}(400+361+\cdots+676)} \approx 12.52$.

3. **5+2+2+3+2=14 points** The accompanying data set consists of observations on shower-flow rate (L/min) for a sample of $n = 30$ houses in Perth, Australia:

4.6	12.3	7.1	7.0	4.0	9.2	6.7	6.9	11.5	5.1
10.5	10.3	8.0	8.8	6.4	5.1	5.6	9.6	5.5	7.5
6.2	5.8	2.3	3.4	10.4	9.8	6.6	3.7	6.4	8.3
6.5	7.6	2.3	13.2	7.3	5.0	6.3	11.8	6.2	5.4
4.8	7.5	6.0	6.9	10.8	7.5	6.6	5.0	3.3	16.2

(a) Construct a stem-and-leaf display of the data.

(b) What is a typical, or representative, flow rate?

(c) Does the display appear to be highly concentrated or spread out?

(d) Does the distribution of values appear to be reasonably symmetric? If not, how would you describe the departure from symmetry?

(e) Would you describe any observation as being far from the rest of the data (an outlier)?

Solution:

(a)

ones digit	the first digit after the decimal point
2	3 3
3	3 4 7
4	0 6 8
5	0 0 1 1 4 5 6 8
6	0 2 2 3 4 4 5 6 6 7 9 9
7	0 1 3 5 5 5 6
8	0 3 8
9	2 6 8
10	3 4 5 8
11	5 8
12	3
13	2
14	
15	
16	2

(b) Open-ended question. Any reasonable answers.

Some possible answers: * mean = 7.216

* median = 6.65

(c) Open-ended question. Any reasonable answers.

Can say "It concentrates around 6".

(d) Open-ended question.

For example, we can say it's skewed to the right, but the main bulk of the data seems to be symmetric etc.

(e) 16.2.

4. **10+2=12 points** The following are the ages of students whose school grades range from kindergarten to senior year of high school ($n = 50$ data points):

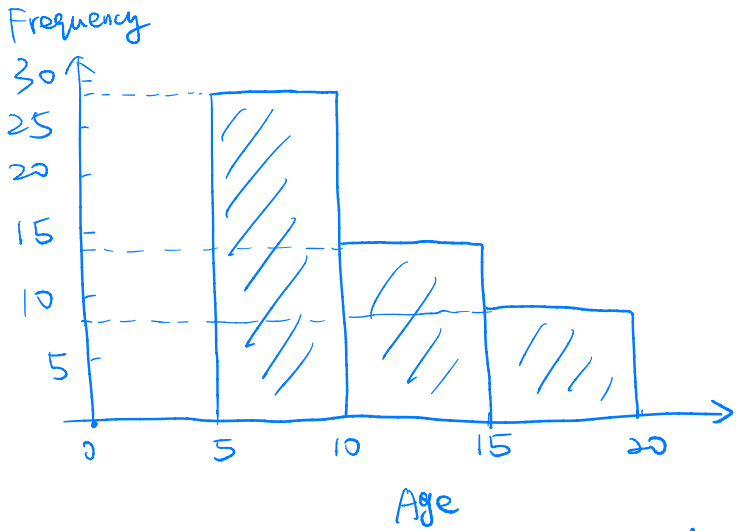
   12, 10, 10, 10, 15, 15, 12, 5, 10, 15, 8, 18, 10, 16, 16, 18, 5, 7, 10, 15, 14, 18, 6, 7, 8, 15, 9, 6, 11, 13, 9, 12, 18, 6, 8, 6, 7, 8, 17, 6, 10, 6, 15, 14, 6, 6, 12, 18, 10, 9

   (a) Construct a histogram with bin-width = 5 and right closed (any one histogram — the frequency one, relative frequency one or density one, as you like).

   (b) How would the shape of the histogram change if it were drawn right open? Provide at least one example of a bar that would be a different height. (You don't have to actually draw it.)

Solution: (a) $n = 50$. Frequency table:

Intervals	Counts (Frequencies)
$[5, 10]$	28
$(10, 15]$	14
$(15, 20]$	8



(b) There are some "10"'s and "15"'s, so the first & second bars will be affected. For example, the height of the first bar will be lower (since some "10"'s will be included by the 2nd bar instead). □

5. **10+2+2+2=16 points** The numbers of children of two samples of mothers is shown below. One sample of mothers had fewer years of education than the other sample (six years or fewer for mothers in the first sample, and seven years or more for those in the other sample).

   Mother educated for six or fewer years:

   0, 0, 2, 2, 3, 4, 5, 5, 9, 10, 10, 11, 13, 13, 13, 13, 14, 14, 14

   Mother educated for seven years or more:

   0, 0, 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 7, 9, 9, 10, 10, 16

   (a) Draw boxplots for each of the samples.

   (b) How do the boxplots differ?

   (c) Are there any outliers? If there are, demonstrate mathematically.

   (d) What can you conclude from the boxplots you drew in part (b) about family size and mothers' education? In your explanation, describe one aspect of shape/spread of the data, and what this means in the context of the question.

Solution: (a) ① ≤ 6 years: $Q1 = 3.5$, $Q2 = 10$, $Q3 = 13$

$IQR = Q3 - Q1 = 9.5$

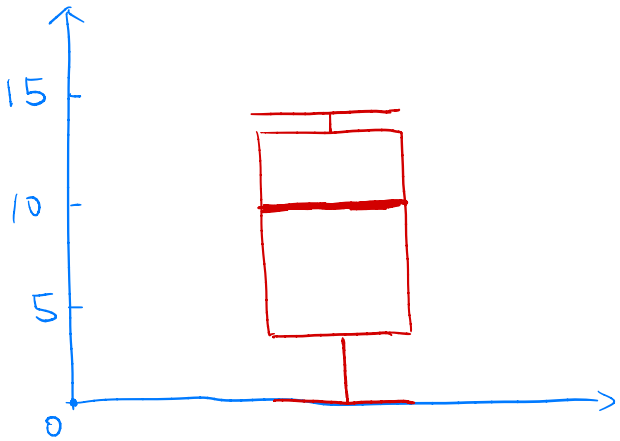
Max upper whisker reach $= Q3 + 1.5IQR = 27.25$

Max lower whisker reach $= Q1 - 1.5IQR = -10.75$

$\text{Max} = 14 < 27.25 \Rightarrow$ upper whisker $= 14$

$\text{Min} = 0 > -10.75 \Rightarrow$ lower whisker $= 0$.

No outliers.



② ≥7 years: $Q_1 = 2$, $Q_2 = 4$, $Q_3 = 5.5$

$IQR = Q_3 - Q_1 = 3.5$

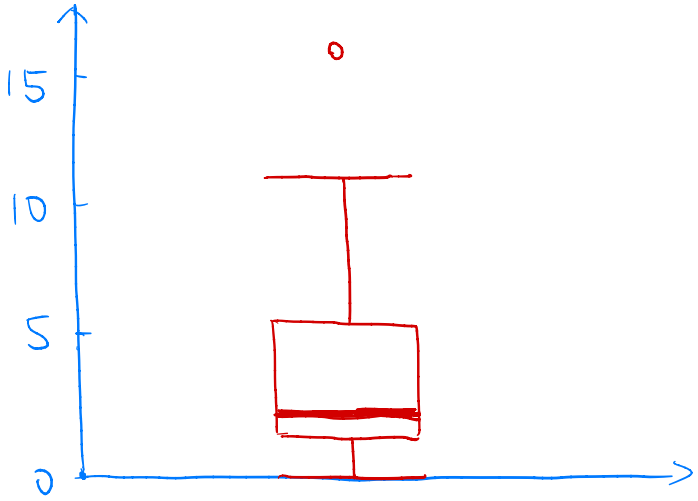
Max upper whisker reach $= Q_3 + 1.5IQR = 10.75$

Max lower whisker reach $= Q_1 - 1.5IQR = -3.25$

Max $= 16 > 10.75 \Rightarrow$ upper whisker $= 10.75$

Min $= 0 > -3.25 \Rightarrow$ lower whisker $= 0$

Suspected outlier: 16



(b) Open-ended. E.g.: * The box of ① is longer than that of ②.
* There's one suspected outliers in ② but none for ①.

(c) None for ①, one for ②, i.e. 16 (See 1a).

(d) Open-ended. E.g.: Mothers educated for ≥ 7 yrs tend to have a smaller family size (less children). Compared to the samples with ≤ 6 yrs education, the standard deviation is also smaller and the data is more concentrated. □.

6. **2+2+2+2=8 points** Here are summary statistics for the sizes (in acres) of the Finger Lakes vineyards.

Count	36
Mean	46.50 acres
StdDev	47.76
Median	33.50
IQR	36.50
Min	6
Q1	18.50
Q3	55
Max	250

Suppose you didn't have access to the data. Answer the following questions from the summary statistics alone:

(a) Would you describe this distribution as symmetric or skewed? Explain.

(b) Are there any outliers? Explain.

(c) Create a boxplot of these data.

(d) Write a few sentences about the sizes of the vineyards.

Solution: (a) Right-skewed, because mean > median.

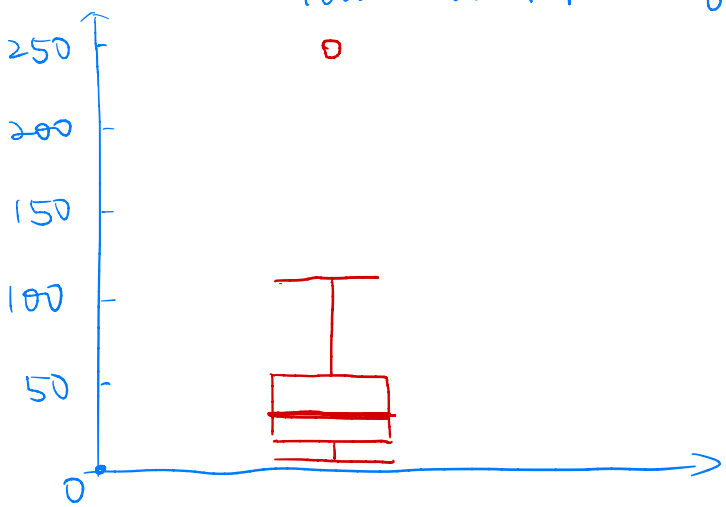
(b) Max upper whisker reach = Q3 + 1.5IQR = 109.75

Max lower whisker reach = Q1 − 1.5IQR = −36.25.

Max = 250 > 109.75 ⇒ 250 is a suspected outlier.

(c) By (b): Upper whisker = 109.75

lower whisker = 6



(d) Open-ended. The distribution is skewed to the right. The middle 50% of vineyards (ie. 25% – 75%) have size from 18.50 acres to 55 acres.

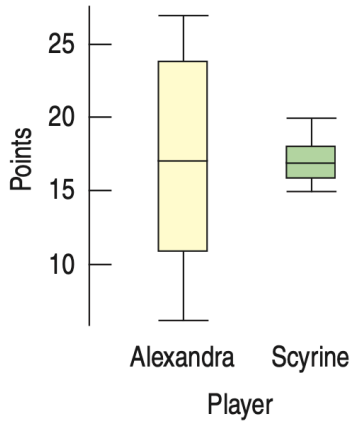
Homework 1

S1201, Summer 2022

100 points

Due July 14, 23:59PM EST

7. **4+4=8 points** Here are boxplots of the points scored during the first 10 games of the season for both Scyrine and Alexandra:



(a) Summarize the similarities and differences in their performance so far.

(b) The coach can take only one player to the state championship. Which one should she take? Why?

Solutions: (a) Open-ended question.
They have similar median.
Alex seems to have a larger variance, while the performance of Scyrine is more stable.

(b) Scyrine, due to more stable performance (less variance). □.

8. **2+2+2+2=8 points** Suppose A and B are some events, and $\mathbb{P}(A) = 0.6$, $\mathbb{P}(B) = 0.4$. Answer the following questions.

   (a) Is it possible for A and B to be mutually exclusive? Explains why.

   (b) Give the maximum and minimum values of $\mathbb{P}(A \cup B)$.

   (c) Give the maximum and minimum values of $\mathbb{P}(A \cap B)$.

   (d) Use Venn diagram to illustrate when these values are achieved in part (b) and part (c).

Solution: (a) Yes.

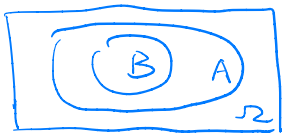
An example: When $A \cup B$ = sample space Ω, A & B can be mutually exclusive. E.g: There are 10 cards with numbers 1 – 10 on them, respectively. We randomly sample one card out. Consider the number on it. $\Omega = \{1, 2, 3, \cdots, 10\}$, $A = \{1, 2, 3, \cdots, 6\}$, $B = \{7, 8, 9, 10\}$.

(b) $A \subseteq A \cup B \subseteq \Omega$

By the property of probability:

$0.6 \leq \mathbb{P}(A) \leq \mathbb{P}(A \cup B) \leq \mathbb{P}(\Omega) = 1$.

Check: 0.6 is achievable:



1 is achievable:

$\Omega = A \cup B$.

(c) $\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) = 1 - \mathbb{P}(A \cup B)$.

By (b): $0 \leq \mathbb{P}(A \cap B) \leq 0.4$.

Check: 0 is achievable:

0.4 is achievable

(d) See (b) & (c). □

9. **4+4+4=12 points** Suppose you have a **fair** coin which can land either H (Heads) or T (Tails) with equal likelihood. We flip this coin 3 times. One possible sequence of outcomes we could get is HHT, if the coin lands Heads the first two times we flip it and tails the 3rd time.

   (a) How many total possible outcomes are there? Please list them all.

   (b) What's the probability that the first time we get an H?

   (c) What's the probability that we see 2 H's and 1 T?

Solution: (a) 8 : HHH, HHT, HTH, HTT, THH, THT, TTH, TTT.

(b) By (a): $|\Omega| = 8$

Denote event $A = \{$ the first time we get an H $\}$.

Then $A = \{HHH, HHT, HTH, HTT\}$

$\Rightarrow P(A) = \frac{n(A)}{|\Omega|} = \frac{4}{8} = \frac{1}{2}$

(c) Denote event $B = \{$ 2 H's & 1 T $\}$

$\Rightarrow B = \{HHT, HTH, THH\}$

$\Rightarrow P(B) = \frac{n(B)}{|\Omega|} = \frac{3}{8} = 0.375$ □

10. **(Extra problem, not required) 5 bonus points** Doctor Strange goes to a casino and plays dice. The players roll two **fair** dice once, and will get \$100 if both two dice land with the 6 face up. Otherwise they pay \$10. Doctor Strange uses his magic to replace numbers 1,2,3 on both dice with 6 before playing. Then what's the probability that he could win \$100?

    Hint: Although the numbers are changed, the likelihood that the die lands with each face up is still the same.

Solution: The sample space:

all equal to (6,6)

Die 1 \ Die 2	6	6	6	4	5	6
6	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
6	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
6	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

equal to (6,6)

The number of outcomes $N = |\Omega| = 36$.

Denote $A = \{$ He wins \$100 $\} = \{$ both dice are 6 $\}$.

Based on the hint, it's easy to see that

A includes 16 outcomes with equal probability

(One way to look at this is to consider which face is up instead of which number is up. Changing the number doesn't change the face. From this perspective, it's still an experiment with equally likely outcomes)

$$\Rightarrow P(A) = \frac{n(A)}{N} = \frac{16}{36} = \frac{4}{9} \approx 0.444 \quad \square.$$