University of Waterloo STAT 220 – W. H. Cherry

Assignment 2

A2 – 1. The data set at the right gives the scores (y_j; out of 1,000) obtained by n = 66 students on an English language proficiency examination.

(a) Prepare a frequency table and histogram (as on the overleaf side of Figure 3.7 of the Course Materials) for these data, using a class width of 50 and starting at: (i) 300; (ii) the lowest score. Comment briefly on the similarities and differences of the two histograms.

(b) For this sample of scores, find the average, median, mode, standard deviation and IQR, given that:

$$\sum_{j=1}^{66} y_j = 34{,}155, \qquad \sum_{j=1}^{66} y_j^2 = 18{,}076{,}491.$$

345	395	563	505	402	472
691	624	523	461	490	530
516	444	604	406	475	523
582	575	439	523	556	479
629	490	730	611	468	574
420	596	585	354	494	439
446	505	585	468	578	603
417	585	593	417	486	604
515	523	545	505	527	384
431	574	494	560	464	549
541	468	629	607	490	549

A2 – 2. (a) An instructor gives a quiz with three questions, each worth one point; 40% of the class score 3 points, 30% score 2 points, 20% score 1 point and 10% score 0.

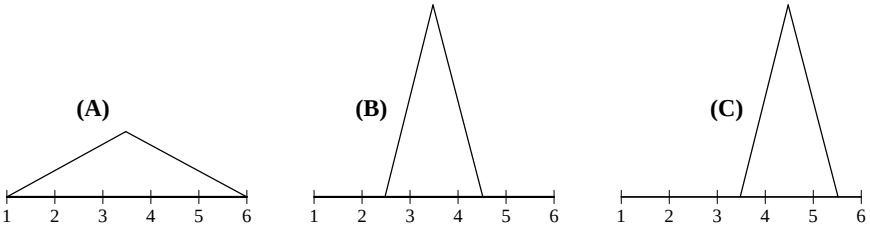
(i) If 10 people wrote the quiz, find the average score.

(ii) If 20 people wrote the quiz, find the average score.

(iii) If the *number* of people who wrote the quiz is not given, explain whether it is still possible to find the average score.

(b) Sketches of the histograms for three data sets are shown below. Match each histogram with *one* of the six average (**Av.**) and standard deviation (**S.d.**) pairs given in the list on the left of the sketches; briefly justify your choice in each case.

	Av.	**S.d.**
(i)	2.5	0.5
(ii)	3.5	0.5
(iii)	3.5	1.0
(iv)	3.5	2.0
(v)	4.5	0.5
(vi)	4.5	1.0



A2 – 3. The distribution of incomes (in dollars) in 1973 tax returns for males aged 25-29 was as shown in the Table at the right; it is known that the average income over $25,000 was $36,210.

(a) Construct a histogram of these data.

(b) Within which income range is the median income? Explain your reasoning.

(c) Find the approximate average income.

(d) Calculate the lower and upper limits of error for your approximate value in (c). [For information, the *actual* average was $9,041.]

(e) Find the approximate s.d. of income; use your value from (c), *not* the actual average, in your calculation.

Income Range	**%**
0 - 2,000	0.6
2,000 - 3,000	2.1
3,000 - 4,000	4.4
4,000 - 5,000	6.1
5,000 - 6,000	7.5
6,000 - 7,000	9.5
7,000 - 8,000	10.7
8,000 - 9,000	12.2
9,000 - 10,000	12.2
10,000 - 15,000	29.2
15,000 - 20,000	4.2
20,000 - 25,000	0.7
25,000 and over	0.6

A2 – 4. The Table at the right is reproduced from a former standard text for nurses.

(a) Construct a histogram of these data using the same classes as in the Table, except for an appropriate half-integer shift in the class boundaries; treat the last (open) intervals '31 and over' as 31 - 45 days.

(b) Find approximate values of the sample average and s.d.

(c) On the basis of these data, how accurate does Naegele's Rule appear to be? Explain your assessment briefly.

(d) From the appearance of the histogram, suggest a probability distribution which might be appropriate as a model for these data.

DEVIATION FROM CALCULATED DATE OF CONFINEMENT, ACCORDING TO NAEGELE'S RULE, OF 4,656 BIRTHS OF MATURE INFANTS*

DEVIATION IN DAYS	**EARLY DELIVERY**	**DELIVERY ON CALCULATED DATE**	**LATE DELIVERY**
0	.....	189 (4.1)†	.....
1- 5	860 (18.5)†	.....	773 (16.6)†
6-10	610 (13.1)	.....	570 (12.2)
11-20	733 (15.7)	.....	459 (9.9)
21-30	211 (4.5)	.....	134 (2.9)
31 and over	65 (1.6)	.....	42 (0.9)

The menstrual cycles of the mothers were 28 ± 5 days. The infants were at least 47 cm in length and 2,600 gm in weight (Burger and Korompai).

*Eastman, N.J.: Williams Obstretics, Ed. 11, p. 216, New York, Appleton, 1956.
†Numbers in parentheses represent per cent of cases considered.

(*continued overleaf*)

A2 – 5. Bird studies are often conducted by capturing and banding birds so that their movements can be followed after they are released. One characteristic of interest is the distance from the point of release to the point of first landing; the two tables at the right give samples of such data (in feet) for two species of bird.

(a) For each data set, find the average, median, standard deviation, variance, interquartile range and range.

(b) An **outlier** in a data set is an observation which is so far removed (in either direction) from the main body of the data that the appropriateness of including it when the data are analyzed is questionable. The *last* observation (1200.0) in the mourning dove data set is very different from the other 24 observations and *may* therefore be an outlier.

(i) To illustrate the effect of such an observation, omit it from the data set and recalculate the average, median, standard deviation, variance, interquartile range and range.

(ii) By comparing the values found for the six summary statistics in (a) and in (b)(i), explain briefly why it is desirable to give more than one measure of location and of variation for a data set.

(iii) Explain briefly whether there appears to be an outlier in the data set for robins.

.....ROBIN.....

128.8	57.2	48.2
160.0	65.2	69.2
192.1	68.9	117.3
163.4	24.7	36.5
186.4	37.4	140.8
156.2	99.7	59.3
70.0	265.0	71.3
10.0	78.7	105.3

MOURNING DOVE

40.0	381.7	358.9
80.0	266.8	13.9
313.9	162.7	165.5
175.7	76.0	317.2
55.5	22.1	300.6
44.7	170.0	197.7
166.7	263.7	288.1
83.4	369.7	102.0
		1200.0

A2 – 6. Figure 1.36 on page 86 of the text shows three p.d.f.s with three points marked A, B and C on each. For each p.d.f., identify the point(s) at which the mean, median and mode fall; explain your reasoning briefly.

A2 – 7. (a) Text Exercise 1.73 (page 86): *The Environmental Protection Agency requires that the exhaust .....*

(b) Text Exercise 1.75 (page 87): *Give an interval that contains the middle 95% of NOX levels .....*

A2 – 8. Text Exercise 1.93 (page 90): *The scores of a reference population on the Wechsler Intelligence Scale .....*

A2 – 9. Master, Dublin and Marks [*J. Amer. Med. Assoc.* **143**: *1464-1470 (1950)*] suggest that a person can be classified as hypotensive or hypertensive if their systolic blood pressure lies, respectively, in the bottom or top 5% of the distribution of blood pressures for their age group. For males aged 20-24 years, these authors found an average systolic pressure of 122.9 mm of mercury and an s.d. of 13.74 mm of mercury. If systolic blood pressure in this age group can be modelled by a normal distribution, find:

(a) the upper limit (u) for the blood pressure of a person classified as hypotensive;

(b) the lower limit (l) for the blood pressure of a person classified as hypertensive.

A2 – 10. (a) Suppose that the weight of tomato juice in machine-filled cans can be modelled by a normal distribution with an s.d. of 8 grams; this variation is a characteristic of the machine that fills the cans. The *average* weight of juice that the machine puts into the cans can be set by an adjustment on the machine. A substantial fine may be incurred if inspection shows that there is less than 454 grams of juice in more than 10% of cans labelled as containing 454 grams. To what average should the machine be set so that *exactly* 10% of the cans will contain less than 454 grams of juice?

(b) Let the random variable R represent the amount of radiation that can be absorbed by an individual before death ensues; assume that R can be modelled by a normal distribution with a mean of 500 roentgens and an s.d. of 150 roentgens. Above what dose (d) of radiation will only 2% of those who receive the dose survive?

(c) Lead, like most other elements, has always been present in the natural environment, but the industrial revolution and the advent of the automobile have increased background environmental lead levels to an extent where some individuals may acquire dangerously high levels of lead in their blood. Let the random variable L represent the blood lead level in μg/dl (micrograms per decilitre), and assume that L can be modelled by a normal distribution with a mean of 25 μg/dl and an s.d. of 11 μg/dl. If a blood lead level of 60 μg/dl or higher is considered to be of serious concern, what proportion of people selected equiprobably from the population will have such high levels?

A2 – 11. The volume of pop placed in 750 ml bottles by a bottling machine can be modelled by a normal distribution with mean μ and standard deviation σ. Over a long period of time, it is observed that 10% of the bottles contain less than 745 ml, while 1% contain more than 765 ml.

(*continued on the overleaf side of Assignment 1*)