University of Waterloo — STAT 220 – W. H. Cherry

Assignment 2 Outline Solution

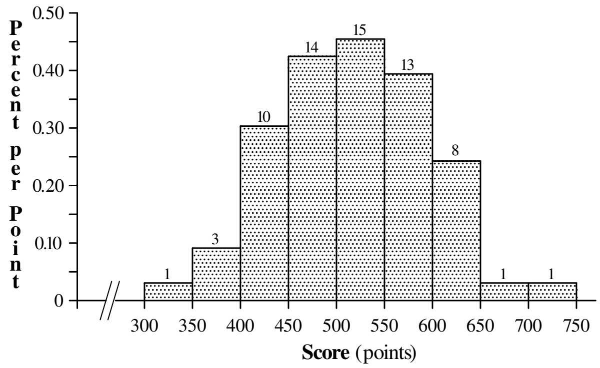
A2 – 1. (a) The ordered data set (or *order statistic*) is given at the right; it leads to the frequency tables and histograms shown below; the latter illustrate how, for some data sets, two reasonable but different choices for the classes lead to histograms that indicate *different* shapes for the distribution of the data (here, unimodal *vs* bimodal).

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

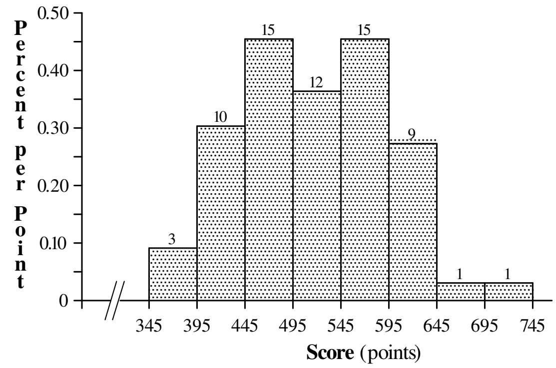
SCORE (points)	**FREQUENCY** No.	%	**DENSITY** (% per point)
[300,350)	1	1.52	0.0303
[350,400)	3	4.55	0.0909
[400,450)	10	15.15	0.3030
[450,500)	14	21.21	0.4242
[500,550)	15	22.73	0.4545
[550,600)	13	19.70	0.3939
[600,650)	8	12.12	0.2424
[650,700)	1	1.52	0.0303
[700,750)	1	1.52	0.0303

SCORE (points)	**FREQUENCY** No.	%	**DENSITY** (% per point)
[345,395)	3	4.55	0.0909
[395,445)	10	15.15	0.3030
[445,495)	15	22.73	0.4545
[495,545)	12	18.18	0.3636
[545,595)	15	22.73	0.4545
[595,645)	9	13.63	0.2727
[645,695)	1	1.52	0.0303
[695,745)	1	1.52	0.0303

English Proficiency Examination Scores



English Proficiency Examination Scores



(b) Using the values given for $\sum_{j=1}^{n} y_j$ and $\sum_{j=1}^{n} y_j^2$ and the order statistic, we find:

$\overline{y} = 517.5$; median = 519.5; mode = 523; s = 78.57 ≃ 79; IQR = 578 – 468 = 110.

A2 – 2. (a) Assume that n people wrote the quiz.

(i) If n = 10: $\overline{y} = (4\times3 + 3\times2 + 2\times1 + 1\times0)/10 = 2.0$.

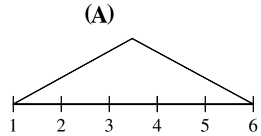
(ii) If n = 20: $\overline{y} = (8\times3 + 6\times2 + 4\times1 + 2\times0)/20 = 2.0$.

(iii) If n people wrote the quiz: $\overline{y} = (0.4n\times3 + 0.3n\times2 + 0.2n\times1 + 0.1n\times0)/n = 2.0$;
i.e., it *is* possible to find the average score without the *value* of n.

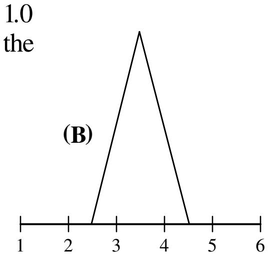
(b) Each histogram is *symmetrical* so that each average is the *centre* of the relevant histogram.
Thinking of the s.d. as the 'average' deviation of the observations of a data set from its average, an s.d. of 0.5 is too small and of 2.0 is too large for (**A**) so that it must be 1.0 of the three available options. Similarly, (**B**) and (**C**) have the *same* s.d. and it is smaller than that of (**A**), making it 0.5 of the options available.

	Av.	**S.d.**
(**A**)	3.5	1.0
(**B**)	3.5	0.5
(**C**)	4.5	0.5

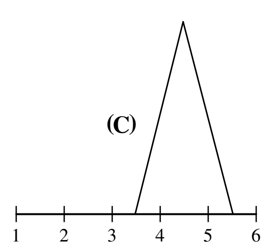
(A)



(B)



(C)



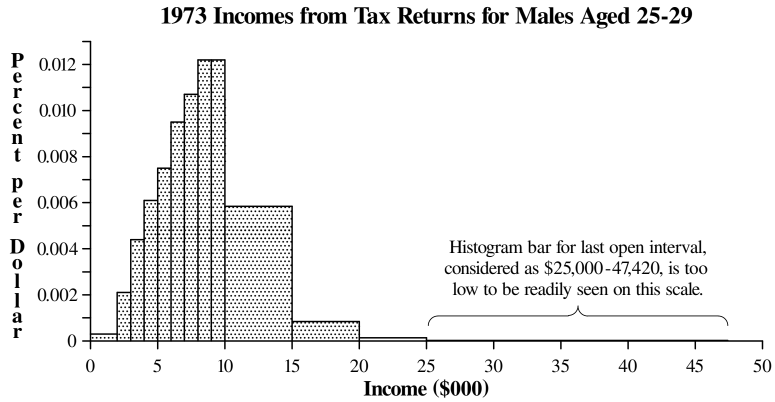
A2 – 3. (a) We use the same classes as in the Table of data – the only difficulty is that we do not know the upper limit of the last open interval; one reasonable option is to make it symmetrical about its average (*i.e.*, regard the interval as $25,000-47,420). The histogram is shown on the following page 0.10b.

(*continued overleaf*)

Assignment 2 Outline Solution (continued 1)

A2 – 3. (cont.) (a)

INCOME ($)	FREQ. %	DENSITY (% per $)
0- 2,000	0.6	0.000 300
2,000- 3,000	2.1	0.002 100
3,000- 4,000	4.4	0.004 400
4,000- 5,000	6.1	0.006 100
5,000- 6,000	7.5	0.007 500
6,000- 7,000	9.5	0.009 500
7,000- 8,000	10.7	0.010 700
8,000- 9,000	12.2	0.012 200
9,000-10,000	12.2	0.012 200
10,000-15,000	29.2	0.005 840
15,000-20,000	4.2	0.000 840
20,000-25,000	0.7	0.000 140
25,000 and over	0.6	0.000 027



(b) Up to an income of $8,000, the percentages in the middle column of the Table above total 40.9%, which is *less than* 50%; up to an income of $9,000, they total 53.1%, which is *more than* 50%. Hence, the median must lie in the interval $8,000-9,000.

(c) To find an approximate value for this average from *grouped* data, as discussed in Section 7 of Figure 4.1 of the Course Materials, we take the sum of the products of the mid-points of the intervals and the percentages, and then divide the sum by 100 (the total of the percentages); the only exception is for the last open interval where we use the *actual average* as given ($36,210) *instead of* the interval mid-point. The result is $9,275.26 ≃ 9,275.

(d) Finding an approximate average in (c) is based on the assumption that, within each class except the last, incomes are *uniformly* distributed over the class interval. The *most* by which this assumption can be wrong is if all the incomes in a class lie at one of its *end-points*. Hence, the *lower* limit for the true average income is obtained from the same type of calculation as in (c) but using the *lower* ends of the 12 closed intervals in place of their mid-points; the result is $8,093.26 ≃ $8,093. The corresponding calculation for the *upper* limit, using the *upper* ends of the 12 closed intervals, yields $10,457.26 ≃ $10,457. Comparing these limits with the *actual* average of $9,041 shows that they are *very* conservative; in fact, the agreement of the approximate value from (c) with the actual average shows that the assumption of uniformly distributed incomes within each interval is reasonable, at least for the data set as a whole.

(e) As discussed in Section 8 of Figure 4.8 of the Course Materials, we need the sum of the *squares* of the observations to find the s.d.; we find an approximate value of this sum from grouped data by means of a calculation similar to that in (c) for the average, except that we use the *squares* of the interval mid-points; the approximate value for this sum is 103,935,984.60 $^{2}. Subtracting the quare of the approximate average from (c) and taking the square root gives an approximate s.d. of $4,231.9 ≃ $4,230.

NOTE: Because the procedure in (e) for the approximate s.d. from grouped data involves the *squares* of the interval mid-points, any error involved in departures from the assumption of uniformly distributed incomes within a class is *squared*; the approximate s.d. from grouped data is therefore generally *less* accurate than the corresponding approximate average.

A2 – 4. (a) When a baby is said to be born on the *calculated* day, we take this to mean anywhere within a specified 24-hour period; thus, we classify the 189 births on the calculated date as falling in an interval of −0.5 to 0.5 days, being the hypothetical calculated delivery time. Since time is a *continuous* quantity, the other interval end-points need to be similarly modified by half a day; mathematically, we regard early deliveries as *negative* and late deliveries as *positive*. The frequency table and histogram are shown on the following page 0.10c.

(b) To find an approximate value for this average from *grouped* data, as discussed in Section 7 of Figure 4.1 of the Course Materials, we take the sum of the products of the mid-points of the intervals and the frequencies and then divide the sum (which is 8,045.5) by 4,656 (the total of the frequencies; *i.e.*, the total number of births). The result is −1.728 ≃ −2 days, or about 2 days *early*.

To find an approximate s.d. from grouped data, as discussed in Section 8 of Figure 4.8 of the Course Materials, we find the approximate sum of the squared observations as 769,879.25. Subtracting $4{,}656 \times (-1.728)^2$, dividing by (4,656 − 1) and taking the square root gives an approximate s.d. of 12.74 ≃ 13 days.

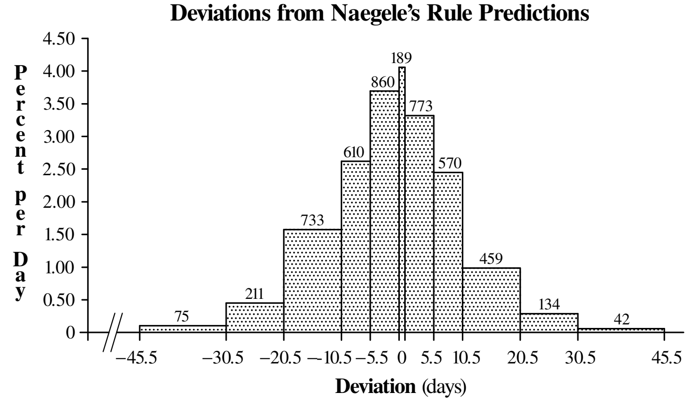
(c) The 'normal' duration of pregnancy is 280 days, so that an average deviation of only about 2 days in the predictions of Naegele's rule over more than 4,650 normal births may *appear* to provide strong evidence that the rule is

Assignment 2 Outline Solution (continued 2)

A2 – 4. (cont.) (c) accurate. However, this relatively small 2-day error is the result of the positive and negative deviations largely cancelling each other and so is misleading; in fact, the original Table shows that there are appreciable numbers of quite large deviations from the expected date of delivery. A more accurate assessment of Naegele's rule is provided by the average *absolute* deviation – we find its approximate value from the grouped data in a manner analogous to the ordinary average in (b), except that we take *all* the interval mid-points as positive. The approximate value for the sum of the absolute deviations is 46,058.5 which, divided by 4,656, gives 9.892 ≃ 10 days; this value shows the rule in a considerably *less* favourable light. Hence, our final answer is that Naegele's rule is of only *moderate* accuracy – deviations from the predicted date of the order of 10 days or greater occur for an appreciable proportion of births.

DEVIATION (days)	FREQUENCY No.	%	DENSITY (% per day)
[−45.5, −30.5)	75	1.61	0.1073
[−30.5, −20.5)	211	4.53	0.4532
[−20.5, −10.5)	733	15.74	1.5743
[−10.5, −5.5)	610	13.10	2.6203
[−5.5, −0.5)	860	18.47	3.6942
[−0.5, 0.5)	189	4.06	4.0593
[0.5, 5.5)	773	16.60	3.3204
[5.5, 10.5)	570	12.24	2.4485
[10.5, 20.5)	459	9.86	0.9858
[20.5, 30.5)	134	2.88	0.2878
[30.5, 45.5)	42	0.90	0.0601

Deviations from Naegele's Rule Predictions



(d) The roughly symmetrical bell-shaped appearance of the histogram suggests a normal distribution with a mean of −1.73 days and an s.d. of 12.74 days as an appropriate probability model for these data.

A2 – 5. (a) The two order statistics are given at the right; also, the summary statistics for the two data sets are:

ROBIN: $\sum_{j=1}^{n} y_j = 2{,}411.6$; $\sum_{j=1}^{n} y_j^2 = 332{,}288.86$ (n = 24).

DOVE: $\sum_{j=1}^{m} z_j = 5{,}616.5$; $\sum_{j=1}^{m} z_j^2 = 2{,}581{,}814.87$ (m = 25).

Using these values and the order statistics, we find the following (in feet; rounded to the nearest integer):

........ROBIN.........			MOURNING DOVE		
10.0	24.7	36.5	13.9	22.1	40.0
37.4	48.2	57.2	44.7	55.5	76.0
59.3	65.2	68.9	80.0	83.4	102.0
69.2	70.0	71.3	162.7	165.5	166.7
78.7	99.7	105.3	170.0	175.7	197.7
117.3	128.8	140.8	263.7	266.8	288.1
156.2	160.0	163.4	300.6	313.9	317.2
186.4	192.1	265.0	358.9	369.7	381.7
					1200.0

	ROBIN	DOVE
Average	100	225
Median	75	170
S.d.	63	235
Variance	3,911	55,000
IQR	90	229
Range	255	1,186

Variance is included in these two tables only for practice calculating it; in the real world, the useful measure of variation is standard deviation.

(Recalculated)	ROBIN	DOVE
(b) Average	93	184
Median	71	168
S.d.	53	120
Variance	2,805	14,308
IQR	84	216
Range	182	368

(b) (i) The values of the *recalculated* numerical data summaries, omitting the largest observation (265.0 and 1,200.0 respectively) from each data set, are given in the Table (b) above at the *right*.

(ii) Comparing the two values of each data summary for the mourning doves in the tables above, we see that the removal of the 'outlier' of 1,200.0 feet causes a substantial reduction in the values of the average, the s.d. and variance, and the range, but has relatively little effect on the two *resistant* summaries, the median and the IQR. Thus, to try to avoid conveying a distorted impression of a data set because of the effects of a small number of 'outliers,' it is good practice to give *both* the average-s.d. *and* the median-IQR pairs of numerical data summaries.

(iii) Comparing the two values of each data summary for the robins in the tables above shows no more change in the values of these summaries than might reasonably be anticipated by removing an observation which is the *largest* in the data set but *is* still part of the data set as a whole; *i.e.*, the observation of 265.0 feet for the robins does *not* appear to be an outlier.

(*continued overleaf*)

Assignment 2 Outline Solution (continued 3)

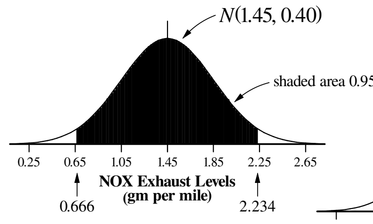
A2 – 6. (a) The longer *right* tail of the distribution *in*creases the mean (C) more than the median (B) and does *not* affect the mode (A).

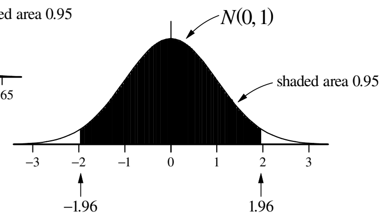
(b) For this *symmetrical* distribution, the mean, median and mode *all* lie at (A).

(c) The longer *left* tail of the distribution *de*creases the mean (A) more than the median (B) and does *not* affect the mode (C).

A2 – 7. (a) The $N(1.45, 0.40)$ p.d.f. is shown at the right.

(b) From Figure 5.3 of the Course Materials, the central 95% of the area of a normal distribution lies within 1.96 s.d.s of the mean; for the $N(1.45, 0.40)$ distribution, this interval is [0.666, 2.234] ≃ [0.67, 2.23] gm per mile.





A2 – 8. (a) The (model) mean is 100 so that 50% of the population is below a score of 100.

(b) A score of 80 is $1.\dot{3}$ s.d.s below the mean; from the $N(0, 1)$ Table (*e.g.*, Figure 5.4 of the Course Materials), the proportion is 9.121 ≃ 9.1%.

(c) A score of 140 is $2.\dot{6}$ s.d.s above the mean; from the $N(0, 1)$ Table, the proportion is 0.3831 ≃ 0.4%.

(d) A score between 80 and 100 is between the mean and $1.\dot{3}$ s.d.s above the mean; from the $N(0, 1)$ Table, the proportion is 40.879 ≃ 40.9%.

A2 – 9. (a) Let the random variable B represent the blood pressure (in mm of mercury) of an equiprobably('randomly')-selected person;
we use the model: $B \sim N(122.9, 13.74)$.

We want: $\Pr(B \le \mathrm{u}) = 0.05$, where u is the required upper limit for a hypotensive.

Hence: $\Pr(\frac{B-\mu}{\sigma} \le \frac{\mathrm{u}-122.9}{13.74}) = \Pr[N(0,1) \le \frac{\mathrm{u}-122.9}{13.74}] = 0.05$,

so that: $\frac{\mathrm{u}-122.9}{13.74} = -1.6449$ which gives: $\mathrm{u} = 100.299 \simeq 100.3$ mm of mercury.

(b) We also want: $\Pr(B \le \mathrm{l}) = 0.95$, where l is the required lower limit for a hypertensive.

Hence: $\Pr(\frac{B-\mu}{\sigma} \le \frac{\mathrm{l}-122.9}{13.74}) = \Pr[N(0,1) \le \frac{\mathrm{l}-122.9}{13.74}] = 0.95$,

so that: $\frac{\mathrm{l}-122.9}{13.74} = 1.6449$ which gives: $\mathrm{l} = 145.501 \simeq 145.5$ mm of mercury.

A2 – 10. (a) Let the random variable W represent the weight of tomato juice (in grams) in a an equiprobably('randomly')-selected can;
we use the model: $W \sim N(\mu, 8)$, where μ is the required machine setting for the average weight of fill.

We want: $\Pr(W < 454) = 0.10$.

Hence: $\Pr(\frac{W-\mu}{\sigma} < \frac{454-\mu}{8}) = \Pr[N(0,1) < \frac{454-\mu}{8}] = 0.10$,

so that: $\frac{454-\mu}{8} = -1.2816$ which gives: $\mu = 464.2528 \simeq 464.25$ gm.

(b) Let the random variable R represent the amount of radiation (in roentgens) absorbed by an equiprobably('randomly')-selected individual;
we use the model $R \sim N(500, 150)$.

We want: $\Pr(R \ge \mathrm{d}) = 0.02$, where d is the required critical dose of radiation.

Hence: $\Pr(\frac{R-\mu}{\sigma} \ge \frac{\mathrm{d}-500}{150}) = \Pr[N(0,1) \ge \frac{\mathrm{d}-500}{150}] = 0.02$,

so that: $\frac{\mathrm{d}-500}{150} = 2.0537$ which gives: $\mathrm{d} = 808.055 \simeq 808$ roentgens.

(*continued*)

Assignment 2 Outline Solution (continued 4)

A2 – 10. (cont.) (c) Let the random variable L represent the blood lead level (in μg/dl) of an equiprobably('randomly')-selected individual; we use the model: $L \sim N(25, 11)$.

We want: $\Pr(L \geq 60)$;

i.e.: $\Pr(\frac{L-\mu}{\sigma} \geq \frac{60-25}{11}) = \Pr[N(0,1) \geq 3.\dot{1}\dot{8}] = 0.000\,7319 \simeq 0.000\,73$;

thus, about 73 people per 100,000 have blood lead levels above 60 μg/dl.

A2 – 11. (a) Let the random variable V represent the volume of pop (in ml) in an equiprobably('randomly')-selected 750 ml bottle; we use the model: $V \sim N(\mu, \sigma)$.

From the information given in the question, we can write:

$\Pr(V < 745) = 0.10$; $\Pr(V > 765) = 0.01$.

Hence: $\Pr(\frac{V-\mu}{\sigma} < \frac{745-\mu}{\sigma}) = 0.10$, $\Pr(\frac{V-\mu}{\sigma} > \frac{765-\mu}{\sigma}) = 0.01$,

so that: $745 - \mu = -1.2816\sigma$; $765 - \mu = 2.3263\sigma$.

Solving these two equations for the unknowns μ and σ gives: $\mu = 752.104 \simeq 752.1$ ml; $\sigma = 5.543\,391 \simeq 5.543$ ml.

(b) We want: $\Pr(V > 760)$;

i.e.: $\Pr(\frac{V-\mu}{\sigma} > \frac{760-752.104}{5.543\,391}) = \Pr[N(0,1) > 1.4244] = 0.07717 \simeq 0.077$;

thus, about 7.7% of the bottles contain more than 760 ml.

A2 – 12. (i) The CVs are: Great Danes: $33.\dot{3} \simeq 33\%$

Chihuahuas: $51.85 \simeq 52\%$

i.e., despite their appreciably *smaller* s.d., the Chihuahuas have the *larger* CV.

(ii) Great Danes are much *larger* dogs than Chihuahuas so that, when their weights are measured in the *same units*, the Great Dane weights will be numbers of appreciably larger magnitude; it would therefore be expected that their variation, as reflected by their s.d., would be larger. Hence, a direct comparison of s.d.s, which does not allow for this difference of magnitude (or *location*), would not necessarily be expected to provide the same information as a comparison of CVs, which are s.d.-to-average *ratios* and so *do* allow for differences of location. The numbers in this question provide a numerical illustration of the matter.

Blank page