**Crown Institute of Higher Education**

BUS104

Week 2 - workshop discussion/practice questions

Table and **Numerical Summary Measures**

The Australian Customs and Border Protection Agency breeds approximately 150 – 200 potential detector dogs each year at its breeding centre in Bulla, Victoria.

In 2012 there were 21 litters bred in the Detector Dog program. The number of surviving pups bred in each litter is provided below.

7, 6, 8, 9, 10, 6, 3, 1, 7, 2, 8, 8, 9, 4, 7, 10, 8, 7, 9, 9, 9

1. Describe as precisely as possible the variable of interest represented by the above data. What units, if any, are applicable to this variable?

2. Are the data categorical or quantitative? If categorical, are the data nominal or ordinal? If quantitative, are the data continuous or discrete (**you should distinguish between “theoretically continuous but discrete in practice” and “simply discrete” and briefly explain the reasoning behind your answer**)? Further, if quantitative, are the data measured on a ratio scale or an interval scale? In relation to your answer to the last question (if appropriate) what are the practical implications of the data being measured on either a ratio or an interval scale (whichever you decided)? How can you most easily decide whether quantitative data are measured on a ratio or an interval scale?

3. If the variable of interest in this situation was represented by the alphabetic character X, and given that the data provided refers to a population (the litter size data for all litters bred at the Bulla facility in 2012), briefly explain the meaning of the following symbols. **Note that you are not required to provide “values” here just explain the meaning of the symbols, however, in your explanations pay particular care to differentiate between (a) and (g), (e) and (h), and, (f) and (i).** .

(a) NX  (b) xi (c) Σxi (d) Σxi2

(e) μx  (f) σx (g) nx (h) x

(i) sx  (j) 

4. Given that for this data;

NX = 21, Σxi = 147 and  = 130,

**Note: You might like to verify these results with the first two being quite straight forward. However, how would you tackle the third one (provide a verbal explanation only – no need to physically check the calculation (unless you are really keen))?**

determine:

(a) The mean of the data.

(b) The range of the data.

(c) The variance of the data.

(d) The standard deviation of the data.

**Note: In recording your answers pay particular attention to the use of appropriate statistical notation (check back to Q3 above), to the specification of appropriate units of measurement and to the rounding of decimal values to an appropriate number of decimal places. With regard to the latter a useful “rule of thumb” when quoting numerical summary measures recorded in the same units as the original data is to specify one more decimal place than that provided in recording the data. In the special case of the variance a useful guide is to round to twice as many decimal places as you would provide in recording the mean of the data.**

5. Provide brief explanations of what the values obtained in Q4(a), (b) and (d) inform you about the data/variable of interest. In your explanation try and use as non-technical language as possible (this is not easy to do especially for Q4(d)) and try and avoid the use of the word “average” in your explanation of Q4(a).

6. Order the raw “pups per litter” data (from smallest to largest) and then use your ordered data set to comment on the existence of the mode (Mo) of the data. **Hint: In your deliberation be aware that the mode is the observation in a data set that occurs with conspicuously highest frequency.** Think carefully about what this definition implies with regard to this data set. Is there an observation in this data set that occurs with conspicuous highest frequency?

7. In Q6 your attention has been drawn to one of the limitations of the use of the mode as a measure of central location/tendency. For quantitative data of a continuous nature there is another common problem, which tends to rule the mode out as a meaningful measure of central location/tendency. Can you describe what this additional problem associated with the determination of the mode of such data might be?

8. Use the percentile location formula, to identify the locations of the 25’th, 50’th, 60’th, 75’th and 80’th percentiles 100 p ) 1+ n (=L p

9. Use the percentile locations identified in Q8 to determine the approximate 25’th, 50’th, 60’th, 75’th and 80’th percentiles. **Note that there are a number of different techniques available for determining approximate percentiles. The method demonstrated to you in lectures (and revisited here in this and the previous question) is the only method accepted in the assessments in this unit. In particular the process used by graphics calculators commonly used in Victorian secondary schools will provide slightly different answers to “our” method and therefore must not be used in this unit.**

10. Briefly explain what the 25’th, 60’th and 80’th percentiles obtained in Q9 tell you about the “pups per litter” data set (**Note that, by the very nature of the percentile calculation that we have introduced you to in lectures, the word “approximately” should feature prominently in your explanations**).

Manually check the validity of your explanations in each of these three cases by performing appropriate frequency counts of the ordered “pups per litter” data (see Q6). **Note: Don’t be overly disappointed with the fact that your manual checks don’t exactly validate your explanations. This is why the word “approximately” is so important in your explanations. Rest assured that for large data sets (where percentiles are usually used in practice), these checks would return very close results.**

What other terms are commonly used in referring to the 25’th percentile, the 50’th percentile (two of them) and the 75’th percentile?

11. Use two results from Q9 to determine the Inter-Quartile Range (IQR) of the data and briefly explain what information this value conveys to you about the data.