**Week-1 Descriptive Statistics Exercises (with Solutions) [Part 1]**

|  |
| --- |
| **1.** Using the suggested classes, the table of frequencies and relative frequencies are as follows.    A histogram can be drawn using either the frequencies or the relative frequencies - here is the histogram obtained using the frequencies.    This shows us the pattern across the classes at the lower end of the time range, there are quite a lot of classes with frequencies of either zero or one. Although unequal class widths could be used with a histogram, this would require an adjustment to the height of the bar (as area corresponds to the class frequency). Not all readers may understand this approach. Another approach would be to have an open ended class such as “12.5 and above”. |
| 2. For the table of frequencies and relative frequencies we can use, for example, classes of width 5, and class intervals 0 to under 5, 5 to under 10, and so on.    A suitable manner to display the distribution of our data is a histogram, as follows.    The distribution has a peak towards the left, and therefore it is skewed to the right, or positively skewed. |
| **3.** It is appropriate to plot the data as a time series plot.     1. The time series plot shows that the data are clearly seasonal, and in particular more energy is consumed in the winter quarter and less in the summer quarter. You may have chosen to plot total consumption per year (summing the four quarters), or to plot the four values for each year as a separate line – both of these are acceptable. 2. To calculate the mean for fuel consumption over the entire period. The measures are = = 19774/24 = 823.92  |  |  | | --- | --- | | Quarter | Mean | | 1 | 945.17 | | 2 | 724.33 | | 3 | 743.67 | | 4 | 882.50 |   This shows that there is a higher consumption than average on quarter 1 and 4 of each year, with less consumption during the summer. |
| **4.** A suitable graph to investigate a relationship between variables, such as the amount spent on advertising and the sales volume is a scatter plot.    From the graph it seems that as advertising expenditure increases so does sales volume. We would, however, want to test the significance of the relationship to be sure. |
| **5.**  **a.** It is helpful to sort the data in (ascending) order:    The central tendency measures for Company A are:   * = = 22/7 = 3.143 * the median is in position *(n* + 1)/2 = 4 (the fourth position), so the median is 2 * the mode is 1, as it appears 3 times whilst value 2 occurs twice and the values 7 and 8 occur only once each.   For Company B:   * = = 60/12 = 5 * the median is in position *(n+l)/2* = 6.5, so the median is the average of the 6th and 7th values, that is median is (5 + 5)/2 = 5; * the mode is 5, as it appear more times than any other.   **b.**  The standard deviation for Company A:   |  |  |  | | --- | --- | --- | | *x* | *x* - | (*x* - )2 | | 1 | -2.143 | 4.592449 | | 1 | -2.143 | 4.592449 | | 1 | -2.143 | 4.592449 | | 2 | -1.143 | 1.306449 | | 2 | -1.143 | 1.306449 | | 7 | 3.857 | 14.87645 | | 8 | 4.857 | 23.59045 | |  |  |  | |  | total | 54.85714 |   s2 = 54.85714/6  s = √(54.85714/6) = 3.024  For Company B:   |  |  |  | | --- | --- | --- | | *x* | *x* - | (*x* - )2 | | 3 | -2 | 4 | | 4 | -1 | 1 | | 4 | -1 | 1 | | 4 | -1 | 1 | | 5 | 0 | 0 | | 5 | 0 | 0 | | 5 | 0 | 0 | | 5 | 0 | 0 | | 6 | 1 | 1 | | 6 | 1 | 1 | | 6 | 1 | 1 | | 7 | 2 | 4 | |  |  |  | |  | total | 14 |   s2 = 14/11  s = √(14/11) = 1.128  **c.**  For Company A: *range* = *max* - *min* = 8 - 1 = 7  For Company B: *range = max* - *min* = 7 - 3 = 4.  To summarise:   |  |  |  | | --- | --- | --- | |  | Company A | Company B | | mean | 3.143 | 5 | | median | 2 | 5 | | mode | 1 | 5 | | sd | 3.024 | 1.128 | | range | 7 | 4 |   Ideally you want some further information before giving advice. One question would be whether there has been any pattern to the computer repairs given to the two companies. If the problems given to one company are more complex or had more often required parts delivered from a manufacturer, for example, then this may have affected the results. Another important question might be whether the company had accurately predicted repair time.  The advice to give depends on the office’s objectives. Company A has a lower mean, but fluctuates much more than Company B. If, for instance, a daily cost was incurred for each day's wait (perhaps for a rental of another machine or through loss of business) then the mean wait is of more importance and Company A should be chosen. Alternatively, if the firm was able to manage its equipment to cover the broken machine as long as the repair time was known, then Company B might be preferred as it shows much less variation, however knowledge of predictability comes in here, as if Company A can predict repair times well variability is less of an issue. |
| 6.   1. Similar to previous question, it is helpful to sort the data in (ascending) order, at least for the determination of median:   A: 110 120 120 170 220 230 240 250 340 350 450  B: 190 200 200 210 220 240 240 250 270  The central tendency measures for A are:   * = = 2600/11 = 236.36 * the median is in position *(n* + 1)/2 = 6, so the median is 230   For Area B:   * = = 2020/9 = 224.44 * the median is in position *(n+1)/2* = 5, so the median is 220;  1. Standard deviation for A is 108.28 and for B is 26.977.  |  |  |  | | --- | --- | --- | | *x* | *x* - | (*x* - )2 | | 110 | -126.36 | 15966.85 | | 120 | -116.36 | 13539.65 | | 120 | -116.36 | 13539.65 | | 170 | -66.36 | 4403.65 | | 220 | -16.36 | 267.6496 | | 230 | -6.36 | 40.4496 | | 240 | 3.64 | 13.2496 | | 250 | 13.64 | 186.0496 | | 340 | 103.64 | 10741.25 | | 350 | 113.64 | 12914.05 | | 450 | 213.64 | 45642.05 | |  | total | 117254.5 |   s2 = 117254.5/10 = 11725.45  s = √(11725.45) = 108.28   |  |  |  | | --- | --- | --- | | *x* | *x* - | (*x* - )2 | | 190 | -34.44 | 1186.11 | | 200 | -24.44 | 597.31 | | 200 | -24.44 | 597.31 | | 210 | -14.44 | 208.51 | | 220 | -4.44 | 19.71 | | 240 | 15.56 | 242.11 | | 240 | 15.56 | 242.11 | | 250 | 25.56 | 653.31 | | 270 | 45.56 | 2075.71 | |  | total | 5822.22 |   s2 = 5822.22/8 = 727.78  s = √(727.78) = 26.977   1. For A: *range* = *max* - *min* = 450 - 110 = 340   For B: *range = max* - *min* = 270 - 190 = 80   1. The definition of volatility is a tendency to change quickly and unpredictably. If the current market price of housing is one area has a higher range and higher standard deviation which is a measurement of dispersion of data, it is then considered to be more volatile. They should go with area B if they want less volatility.   However, it should also be noted that high dispersion also meant that the price could be extremely lower or higher, so they also risked missing out on a good deal in area A. |
| 7.  Since we are given only grouped data, we need to find the midpoint of each class and use these to calculate an approximate mean and approximate variance.    Note that is possible to work in £ for in thousand £ (as here), BUT in the latter case we must take appropriate action at the end of the calculations.  **a.**    **b.**   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **Price (£000)** | | | **frequency** | **midpoint (mp)** | (mp - ) | (*mp* - )2 | (*mp* - )2f | | 150 | to < | 160 | 21 | 155 | -15.698 | 246.427 | 5174.971 | | 160 | to < | 170 | 27 | 165 | -5.698 | 32.467 | 876.615 | | 170 | to < | 180 | 18 | 175 | 4.302 | 18.507 | 333.130 | | 180 | to < | 190 | 11 | 185 | 14.302 | 204.547 | 2250.019 | | 190 | to < | 200 | 6 | 195 | 24.302 | 590.587 | 3543.523 | | 200 | to < | 210 | 3 | 205 | 34.302 | 1176.627 | 3529.882 | |  |  |  |  |  |  |  |  | |  | Totals |  | 86 |  |  | total | 15708.140 |   So our approximation for s is √(15708.140/86) = 13.515, i.e. £13, 515 |
| **8.**  The total of the weekly salaries is £19406, there are 27 values, so the mean from the raw data is £718.7407.  If we use class widths of 50 we get a table like this (although you may have chosen different class limits for the classes)    Note that the mean is quite close to that from the raw data, but that the lost information has made a difference.  If we use class widths of 100 we get a table like this (although again you may have chosen different class limits for the classes)    Again the mean is near enough to that in the raw data to be useable. In this example the mean gets further from that in the raw data as the classes get wider. This is generally the case, but won’t always happen. |