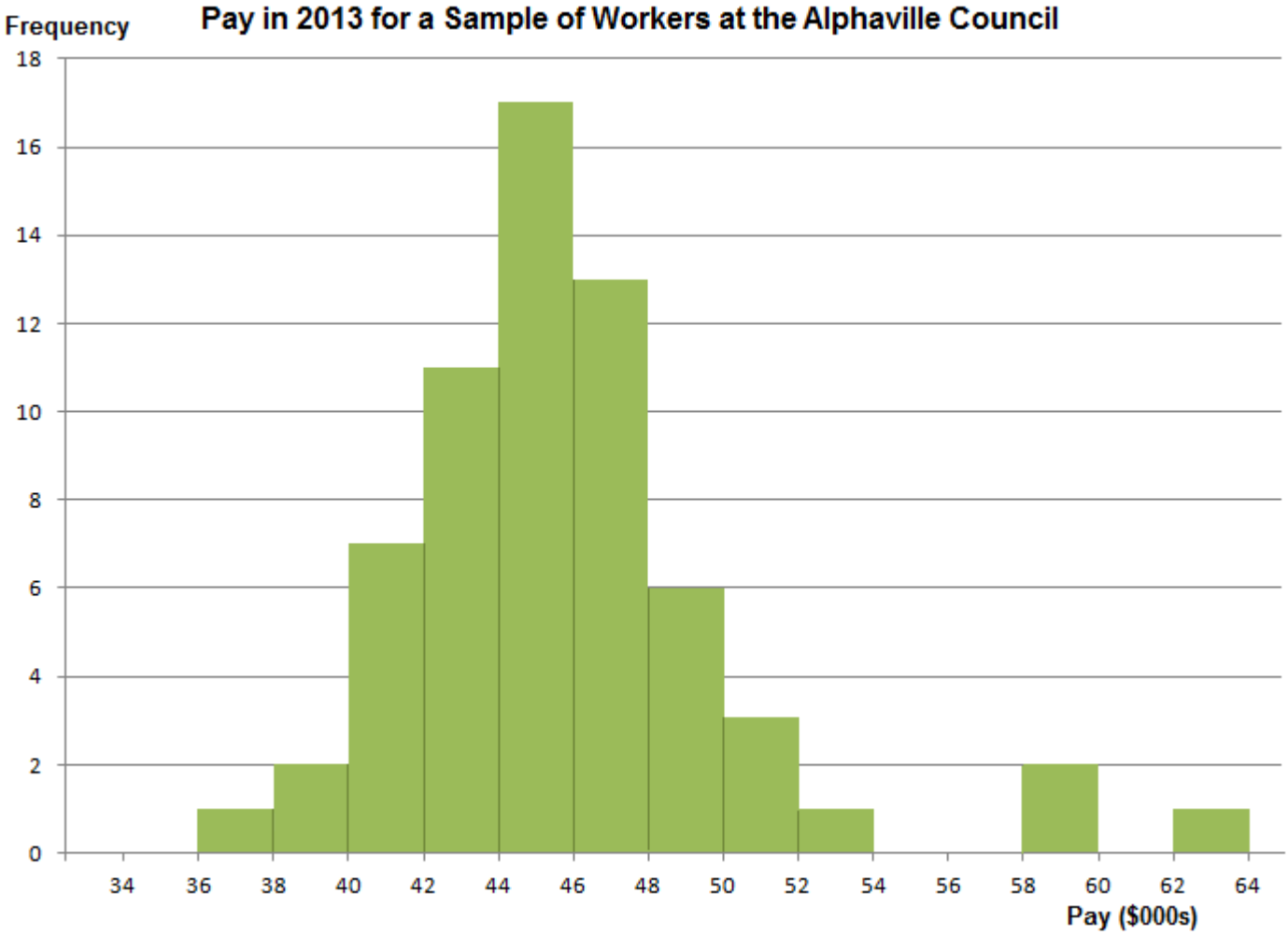
Level 1 Data Practice #7

The pay during 2013 for the 64 workers at the Alphaville Council is shown in the graph below.

The council did not want to show a census of all pay, as that would enable people to work out what individuals were earning.

Eighty people, chosen at random by taking every fifth person on the staff phone list, were asked enter their pay for the year (in $2000 bands) via a secret computer terminal so that the interviewer could not see it. Sixteen refused, and are excluded.

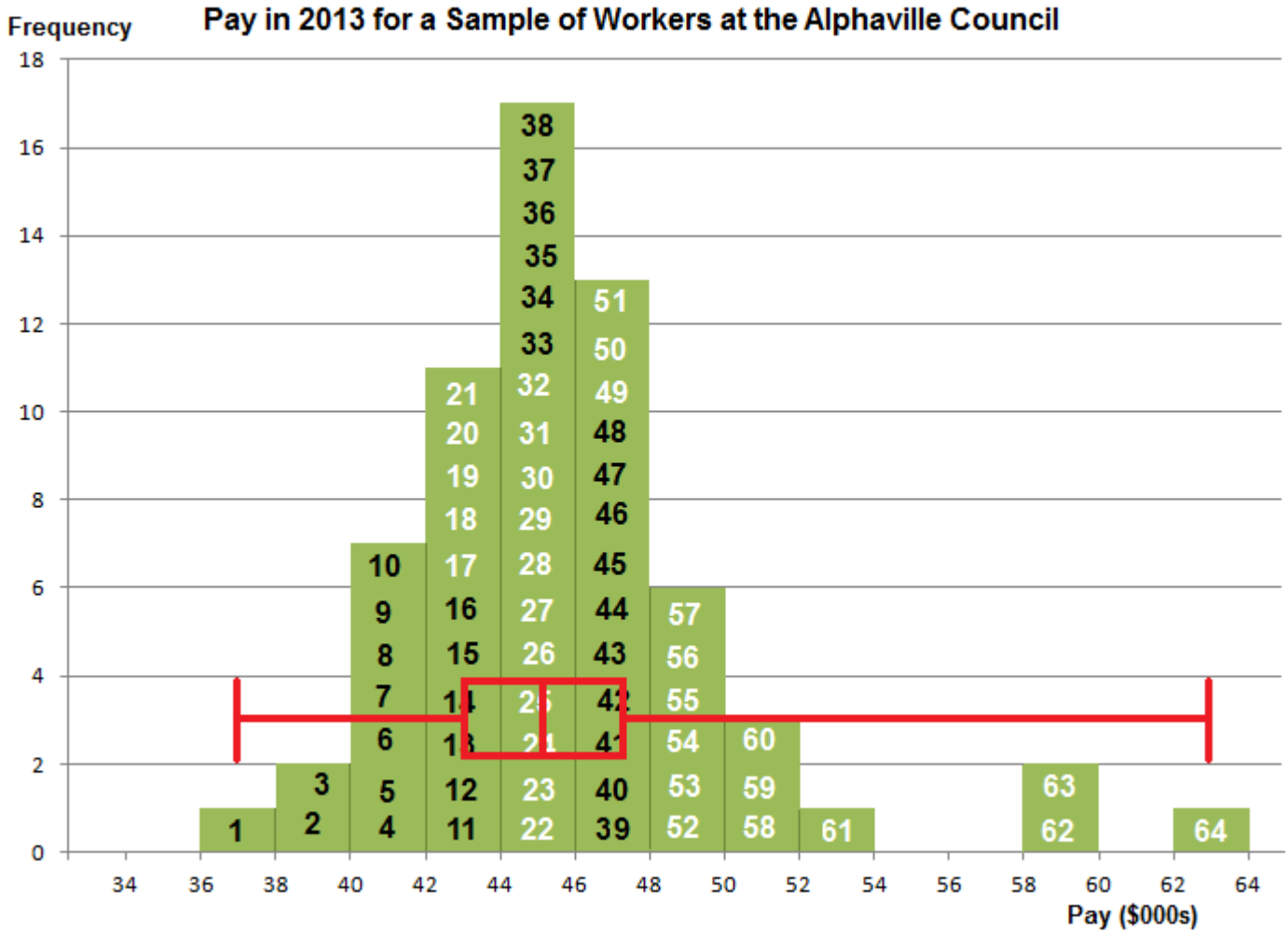


1. Explain a better method of taking an accurate sample of the workers' pay, and why it is better.
2. How many people at the station earned more than $50,000 during 2013?
3. Over the graph draw an estimated box-plot for the same data.

   Explain how you estimated the points for your plot.
4. What do you estimate the mean for the data to be? (Do **not** attempt to calculate it.)

   Explain your estimation.
5. Describe any other features of the distribution.

Answers: Level 1 Data Practice #7



1. The method used is likely to introduce bias, as people will lie about their pay (even when giving it in secret) or not know it accurately. Having many refusals doesn't help it be unbiased. The data should be taken from a pay-roll computer to ensure lack of bias.

   Using every fifth person on the phone list is effective but only if every council worker is on the list – new staff, part-timers, workers outside an office etc must not be left out.

2. Counting the individuals after $50,000 shows that there are 7.

3. The box-plot is drawn in red above.

   The end points had to be estimated at $37,000 and $63,000 as the middle of the end values (the lowest possible of $36,000 and highest of $64,000 are other reasonable choices). The data was divided into quartiles (of 16 each) and the approximate place in each bar was used for the bars of the boxplot. For example the 32nd value is 2/3 of the way into the $44–$46,000 bar, so the median will likely be just more than $45,000.

   (*The important point is to defend choices statistically, exact values are not expected.*)

4. The mean should be about $46,000 or $47,000. It will be slightly more than the median because the data is not symmetrical, and the three extreme values at $58,000 and $62,000 will tend to drag the mean a bit higher than the median.

5. The distribution is unimodal (has a single peak) and symmetrical apart from three much larger values, which give it a skew with a tail at higher pay (the boxplot long whisker).

   (Those three high values are not outliers. They aren't isolated single values and to earn $62,000 a year is not unexpected. Pay is likely to be an unsymmetrical distribution.)