ST104a Live Session Outline Solutions to Exercise 1

1. (a) A stem-and-leaf diagram for the data is:

Stem-and-leaf diagram of speed of cars

Stem (in mph)	Leaves (in 0.1 mph)
25	67788
26	29
27	57899
28	345899
29	0123578
30	112
31	
32	
33	2
34	9

For stem-and-leaf diagrams, in an examination question you would be awarded marks for:

* an informative title
* stem/leaf labels, including units (if provided)
* sensible stems
* vertical alignment
* ordered leaves
* accuracy.

(b) It would be sensible to use the stems from (a) as the classes (bins) for the histogram.

Class interval	Interval width	Frequency	Frequency density
[25.0, 26.0)	1	5	5
[26.0, 27.0)	1	2	2
[27.0, 28.0)	1	5	5
[28.0, 29.0)	1	6	6
[29.0, 30.0)	1	7	7
[30.0, 31.0)	1	3	3
[31.0, 32.0)	1	0	0
[32.0, 33.0)	1	0	0
[33.0, 34.0)	1	1	1
[34.0, 35.0)	1	1	1

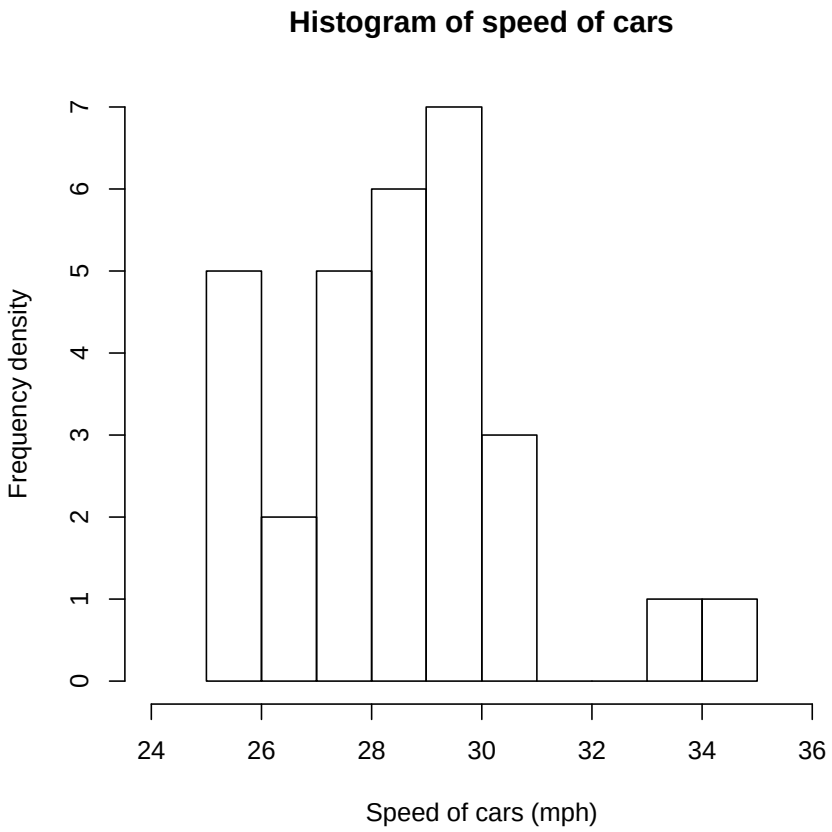
For histograms, in an examination question you would be awarded marks for:

* an informative title
* a 'Frequency density' axis label
* an x-axis label

∗ a sensible number of classes

∗ plotting of frequency densities

∗ accuracy.



(c) The stem-and-leaf diagram and histogram appear to show that the data are positively-skewed (skewed to the right), due to the outlier speeds of 33.2 mph and 34.9 mph.

Note if you are ever asked to comment on the shape of a distribution, consider the following points.

- Is the distribution (roughly) symmetric?
- Is the distribution bimodal?
- Is the distribution skewed (an elongated tail in one direction)? If so, what is the direction of the skewness?
- Are there any outliers?

(d) There are $n = 30$ observations, so the median is the average of the 15th and 16th ordered observations. Using the stem-and-leaf diagram in (a), we see that $x_{(15)} = 28.5$ and $x_{(16)} = 28.8$. Therefore, the median is $(28.5 + 28.8)/2 = 28.65$ mph.

(e) Since Q_2 is the median, which is 28.65, we now need the first and third quartiles, Q_1 and Q_3, respectively. There are several methods for determining the quartiles, and any reasonable approach would be acceptable in an examination. For simplicity, here we will use the following:

$$Q_1 \approx x_{(n/4)} = x_{(7.5)} \approx \frac{x_{(7)} + x_{(8)}}{2} = \frac{26.9 + 27.5}{2} = 27.2 \text{ mph}$$

and:

$$Q_3 \approx x_{(3n/4)} = x_{(22.5)} \approx \frac{x_{(22)} + x_{(23)}}{2} = \frac{29.3 + 29.5}{2} = 29.4 \text{ mph}.$$

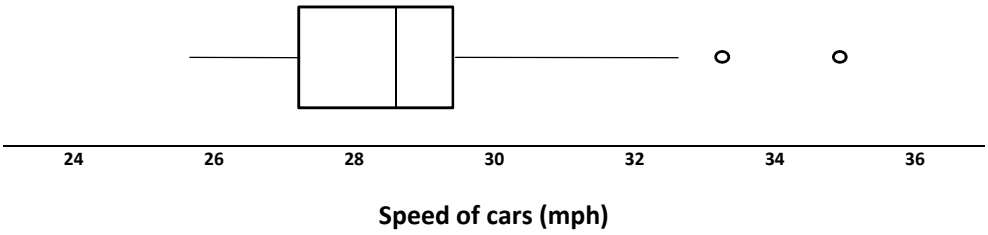
Hence the interquartile range (IQR) is $Q_3 - Q_1 = 29.4 - 27.2 = 2.2$ mph. Therefore, the whisker limits must satisfy:

$$\max(x_{(1)},\, Q_1 - 1.5 \times \text{IQR}) \quad \text{and} \quad \min(x_{(n)},\, Q_3 + 1.5 \times \text{IQR})$$

which is:

$$\max(25.6,\, 23.9) = 25.6 \quad \text{and} \quad \min(34.9,\, 32.7) = 32.7.$$

So there are two observations which lie outside the interval [25.6, 32.7], which are $x_{(29)} = 33.2$ mph and $x_{(30)} = 34.9$ mph and hence these are plotted individually in the boxplot. Since $x_{(29)}$ and $x_{(30)}$ are less than $Q_3 + 3 \times \text{IQR} = 29.4 + 3 \times 2.2 = 36$ mph, then these observations are outliers, rather than extreme outliers. The boxplot is (a vertical orientation is also fine):



(f) We have sample data, not population data, hence the (sample) mean is denoted by $\bar{x}$ and the (sample) standard deviation is denoted by s. We have:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{856.4}{30} = 28.55 \text{ mph}$$

and:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{29}(24{,}574.06 - 30 \times (28.55)^2) = 4.3688.$$

Therefore, the standard deviation is $s = \sqrt{4.3688} = 2.09$ mph.

(g) In (c) it was concluded that the stem-and-leaf diagram and histogram of the data were positively-skewed, although here the mean is (slightly) less than the median. In fact, skewness is not directly related to the relationship between the mean and median: a distribution with negative skew can have its mean greater than or less than the median, and likewise for positive skew. However, this is quite rare. It is possible to quantify skewness, although this is beyond the scope of the course.

(h) We calculate:

$$\bar{x} - s = 28.55 - 2.09 = 26.46 \quad \text{and} \quad \bar{x} + s = 28.55 + 2.09 = 30.64$$

also:

$$\bar{x} - 2 \times s = 28.55 - 2 \times 2.09 = 24.37 \quad \text{and} \quad \bar{x} + 2 \times s = 28.55 + 2 \times 2.09 = 32.73.$$

Now we use the stem-and-leaf diagram to see that 22 observations are between 26.46 and 30.64 (i.e. the interval [26.46, 30.64]), and 28 observations are between 24.37 and 32.73 (i.e. the interval [24.37, 32.73]). So the proportion (or percentage) of the data in each interval, respectively, is:

$$\frac{22}{30} = 0.733 = 73.3\% \quad \text{and} \quad \frac{28}{30} = 0.933 = 93.3\%.$$

Some general points to note:

- Many 'bell-shaped' distributions we meet – that is, distributions which look a bit like the normal distribution – have the property that 68% of the data lie within *approximately* one standard deviation of the mean, and 95% of the data lie within *approximately* two standard deviations of the mean. The percentages in (h) are similar to these.

- When constructing a histogram, it is possible to 'lose' a pattern in the data – for example, an approximate bell shape – through two common errors:

  - too few class intervals (which is the same as too wide class intervals)
  - too many class intervals (which is the same as too narrow class intervals).

  For example, with too many class intervals, you mainly get 0, 1 or 2 items per class, so any (true) peak is hidden by the subdivisions which you have used.

- The best number of (equal-sized) class intervals depends on the sample size. For large samples, many class intervals will not lose the pattern, while for small samples they will. However, with the datasets which tend to crop up in this course, something like 6, 7, 8 or 9 class intervals are likely to work well. So do not forget this!