Solution to Exercise 2.1 (Version 1, 16/09/14)

from Statistical Methods in Biology: Design & Analysis of Experiments and Regression (2014) S.J. Welham, S.A. Gezan, S.J. Clark & A. Mead. Chapman & Hall/CRC Press, Boca Raton, Florida. ISBN: 978-1-4398-0878-8

Exercise 2.1*

A plant ecologist is interested in the distribution of one species of grass within a field. She investigates this by throwing a 0.1 m^2 quadrat to 20 random positions in the field and counting the number of plants of the species in the quadrat at each position. The counts for the 20 quadrats were: 15, 12, 6, 7, 4, 2, 10, 14, 3, 6, 9, 9, 2, 11, 10, 3, 2, 11, 9 and 10. File GRASS.DAT contains the unit number (variate *Quadrat*) and plant count (variate *Count*) for each quadrat. Consider whether these data should be considered as continuous or discrete, and draw a bar chart or histogram (as appropriate). Obtain the sample mean, median and inter-quartile range. What can you say about the distribution of these data?

Solution 2.1

As the data are integer values, the data are discrete and so a bar chart is appropriate and is shown in Figure S2.1.1. (Although if the range of counts was much larger, 0-100 say, we might consider it as approximately continuous and use a histogram with grouped bins). The heights of the bars represent the number of quadrats with each number of plants, e.g. the left-most bar indicates that three quadrats contained two plants. We can see that there were no quadrats with zero, one, five, eight or 13 plants, that the minimum number of plants per quadrat was 2 and the maximum number of plants was 15.

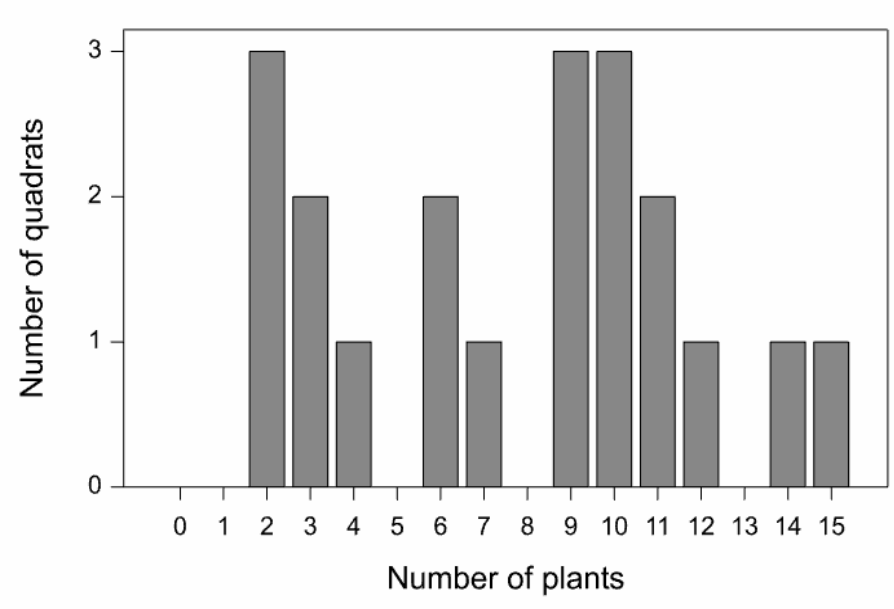

Figure S2.1.1. Bar chart for counts of plants of a grass species in 20 quadrats.

This data set consists of $N = 20$ observations. We denote the i^{th} response as y_i, $i = 1 \ldots 20$. The sample mean, denoted $\bar{y}$, is calculated as the sum of all the observations divided by the sample size, i.e.

$$\begin{aligned}\bar{y} &= \frac{1}{N}\sum_{i=1}^{N} y_i \\ &= \frac{(15+12+6+7+4+2+10+14+3+6+9+9+2+11+10+3+2+11+9+10)}{20} \\ &= \frac{155}{20} = 7.75.\end{aligned}$$

To compute the sample median (or 50^{th} sample percentile) we first compute

$$m = (N + 1) \times 50/100 = (20 + 1) \times 50/100 = 10.5.$$

As m is not integer but lies between 10 and 11, the median is defined as the average of the 10^{th} and 11^{th} values in the ordered set. Ordering the observations (with 10^{th} and 11^{th} underlined) gives

2, 2, 2, 3, 3, 4, 6, 6, 7, <u>9</u>, <u>9</u>, 9, 10, 10, 10, 11, 11, 12, 14, 15.

The 10^{th} and 11^{th} values are both equal to 9 and hence the sample median is 9. Similarly, the sample lower (25^{th}) percentile (lower quartile) is the average of the 5^{th} and 6^{th} ordered observations (as $m = (20 + 1) \times 25/100 = 5.25$), i.e. 3.5, and the sample upper (75^{th}) percentile (upper quartile) is the average of the 15^{th} and 16^{th} ordered observations (as $m = (20 + 1) \times 75/100 = 15.75$), i.e. 10.5. The inter-quartile range is therefore $10.5 – 3.5 = 7$.

From such a small number of observations, it is difficult to draw firm conclusions about the distribution of plant counts. We have found that the median is larger than the mean and, from the bar chart, can see that the distribution of plant counts is not symmetric. For a symmetric distribution, we would expect the counts to be lower at both extremes of the range; here there are some quite high counts for low numbers of plants and quite low counts for intermediate numbers. We might speculate that the distribution is perhaps multi-modal (with some areas of the field having high densities and some much lower) but would need to do further sampling to draw any firm conclusions.

<u>Additional information</u>

The median and the upper and lower quartiles can be illustrated in a **box-and-whisker plot** as shown in Figure S2.1.2. The ‘box’ spans the inter-quartile range, the median is denoted by the solid line across the box, and the ‘whiskers’ extend from edges of the box to the maximum and minimum values. The position of the mean is shown here using a dashed line superimposed on the box-and-whisker plot. For these data the median line is nearer the upper quartile than the lower quartile and the whiskers are of different lengths. For a symmetric distribution the median would be in the centre of the box and the whiskers would be of equal length.

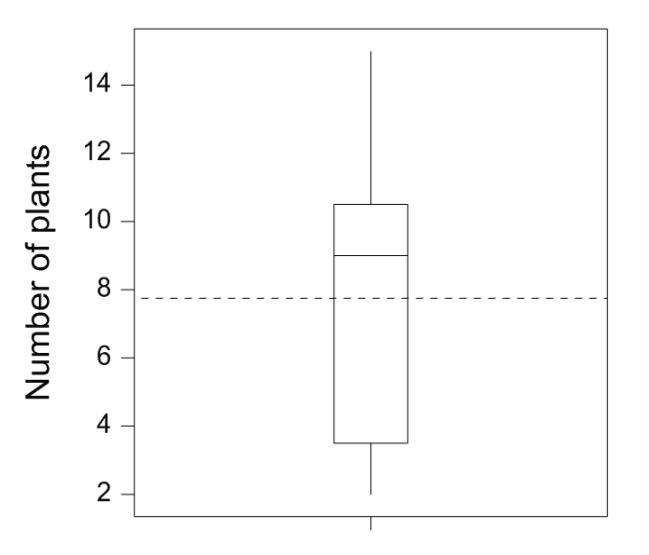

Figure S2.1.2. Box-and-whisker plot for counts of plants of a grass species in 20 quadrats.