Exploration 6.1: Cancer Pamphlets

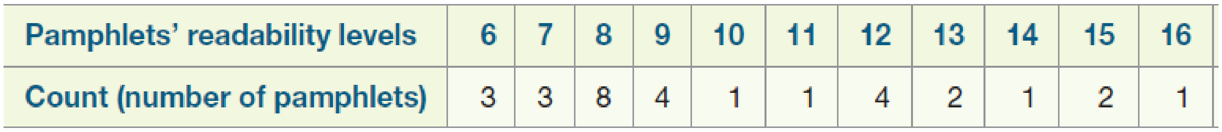
Comparing Two Groups: Quantitative Response

LEARNING GOALS

- Calculate or estimate the mean, median, quartiles, five number summary, and interquartile range from a dataset and understand what these are measuring.
- When comparing two quantitative distributions, identify which has the larger mean, median, standard deviation, and inter-quartile range.
- Identify whether there is likely an association between a binary categorical explanatory variable and a quantitative response variable.

Researchers in Philadelphia investigated whether pamphlets containing information for cancer patients are written at a level that the cancer patients can comprehend (Short, Moriarty, and Cooley, 1995). First, they measured the readability of a sample of cancer pamphlets based on factors such as the length of sentences and the number of polysyllabic words, assigning each pamphlet a grade level. The results shown in Table 6.1.1 are presented as a *frequency table,* reporting the number of pamphlets at each grade level.

Table 6.1.1: Readability measures (grade level) for pamphlets aimed at cancer patients



Pamphlets' readability levels	6	7	8	9	10	11	12	13	14	15	16
Count (number of pamphlets)	3	3	8	4	1	1	4	2	1	2	1

1. What are the observational units in Table 6.1.1? How many observational units were measured?
2. Use the information in Table 6.1.1 to construct a histogram of the pamphlets' readability levels. How would you summarize the behavior of this histogram?
3. Using the information in Table 6.1.1, calculate the mean pamphlet readability level. (*Hint*: Add up the values, 6 + 6 + 6 + 7 + … + 15 + 15 + 16 and divide by the number of observational units.)
4. Using the information in Table 6.1.1, calculate the median pamphlet readability level. How many observations are on each side of this median value (including the repeat values)? (*Hint*: What is the position of the median?)
5. How do the mean and median compare? Is this what you would have predicted from the histogram? Explain briefly.

The mean and median tell us about the center of the distribution. We can also summarize the behavior of the *distribution* of a quantitative variable by dividing the distribution into four pieces of roughly equal size (number of observations). In other words, we can summarize the distribution by indicating where the bottom 25% of the data are, the next 25%, the next 25%, and then the top 25%.

Definition

The value for which 25% of the data lie below that value is called the ***lower quartile (or 25th percentile)***. Similarly, the value for which 25% of the data lie above that value is called the ***upper quartile (or 75th percentile)***. Quartiles can be calculated by determining the median of the values above/below the location of the overall median. The difference between the quartiles is called the ***inter-quartile range (IQR)***, another measure of variability of the distribution along with standard deviation.

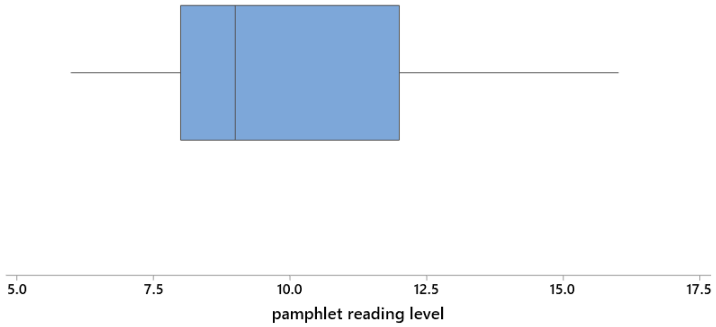
6. Consider the 15 values below the position of the median, and find the median of those 15 values. This is the lower quartile.
7. Repeat for the 15 values above the position of the median. This is the upper quartile.
8. Calculate and interpret the inter-quartile range. (*Hint*: What is the IQR the width of?)
9. Explain why the inter-quartile range might be preferred to the standard deviation to summarize the variability in the pamphlet reading levels.

Key Idea

The IQR is a resistant measure of variability, whereas the standard deviation is sensitive to extreme values and skewness.

The min, lower quartile, median, upper quartile, and max comprise the ***five-number summary***. A visual representation of the five-number summary is a boxplot. Figure 6.1 shows a boxplot for these data.

Figure 6.1.3: Boxplot of cancer pamphlet reading levels



Definition

A boxplot is a visual display of the 5-number summary. The box displays the middle 50% of the distribution, and its width (the IQR) helps us visualize the spread of the distribution; the whiskers extend to the smallest and largest values in the dataset.

10. How does the boxplot match up to the five-number summary?

Endpoint of lower "whisker":

Lower edge of box:

Line within the box:

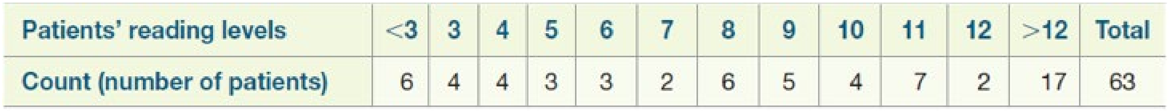
[Type here]

Upper edge of box:

Endpoint of upper whisker:

The above boxplot illustrates that there is some variability in the pamphlet reading levels. What the researchers wanted to know was how these different reading levels matched up to the reading abilities of the cancer patients. Table 6.1.2 shows a frequency table for the reading level (again a grade level) for a sample of 63 cancer patients. Note that patient reading levels of under 3rd grade and above 12th grade are not determined exactly.

Table 6.1.2: Comparing of pamphlet readability and patient reading levels



Patients' reading levels	<3	3	4	5	6	7	8	9	10	11	12	>12	Total
Count (number of patients)	6	4	4	3	3	2	6	5	4	7	2	17	63

11. Explain why it is not possible to calculate the mean reading ability for these patients.
12. Explain why it is possible to calculate the median reading ability for these patients and do so.
13. How does the median reading level of the patients compare to the median reading level of the pamphlets? Does this indicate that the pamphlets are a good match to the cancer patients? Explain.
14. Determine and interpret the lower quartile for the patient reading levels.
15. How does the lower quartile of patient reading levels compare to the lower quartile of the reading level of the cancer pamphlets? How does the lower quartile of patient reading levels compare to the minimum reading level of the cancer pamphlets?
16. Use your answers to question #15 to decide whether these cancer pamphlets are a good match to these patient reading levels. (*Hint*: Interpret the second comparison in the previous question in context.)

Later in this chapter, you will learn formal methods for assessing whether the centers of two distributions are statistically significantly different. But notice how such analysis may not be valid for a study like this:

- Means are not the most important feature to be comparing. For example, means will be influenced by outliers and these data were not presented in a way to allow for calculation of means.
- We could focus on comparing the medians instead, but only comparing the centers of these distributions ignores the variability in the distributions, which is perhaps of more interest to these research question.
- We don't know whether the pamphlets or patients were randomly sampled from larger populations.

This exploration reveals that measures of center do not always tell the whole story when you are analyzing data to address a particular research question. In this case, the research question of whether pamphlets' readability levels are well-aligned with patients' reading levels requires looking at the entire distributions, not simply at measures of center. Examining graphs of the distributions would be a better place to begin. In addition to dotplots and histograms, ***boxplots*** are another potential display that is especially useful for comparing distributions. But keep in mind that boxplots can sometimes mask important features of distributions. Many software packages now allow you to overlay a boxplot on top of a histogram or dotplot to help see the entire distribution while highlighting key features like the ***quartiles***.