How to Interpret Boxplots

While we have seen and plotted “boxplots” for so many times in homework and labs, have you ever wondered what this special kind of plot really tells us? Do you know what the lines, the rectangles, and the dots mean? Most importantly, when are boxplots useful?

In previous introductory statistics courses, all I have learnt was how to create a boxplot from a dataset. I never thought this kind of plot would be useful in any ways, as it seldom appeared outside of the courses. However, when trying to answer the question “Provide concise descriptions for the boxplot.” in homework 1 about the NBA dataset, I realized that a boxplot can contain important information as long as one knows how to read it correctly. I would like to share what I have found from further research on boxplots.

Creating a Boxplot

To begin with, let’s see how to create a boxplot from a sample data.

Say there are 30 students in Discussion 101 of Stat133, and their midterm scores form a vector *midterm_101* :

```
midterm_101 <- c(79, 80, 81, 82, 82, 82, 84, 85, 85, 86, 86, 87, 88, 88, 88,
                 88, 88, 89, 89, 89, 89, 89, 89, 89, 90, 92, 92, 93, 95, 95)
```

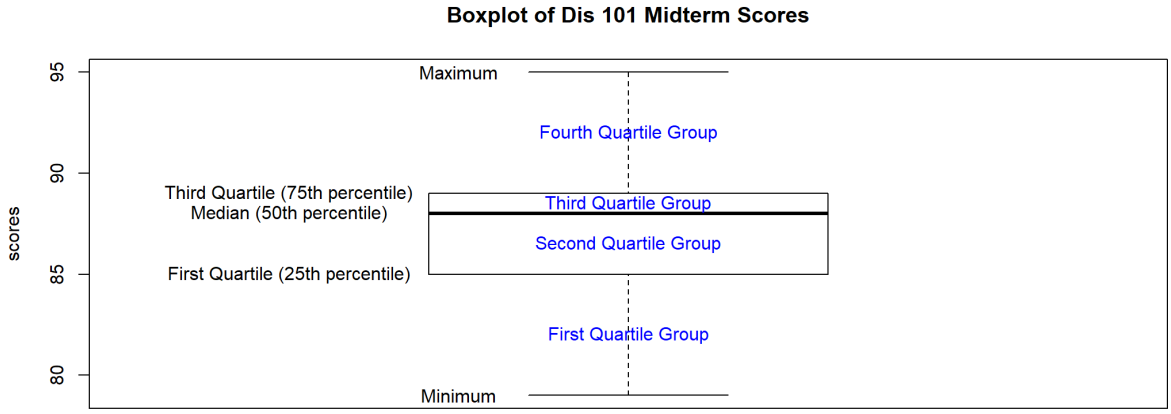
Here is the summary statistics of *midterm_101*:

```
summary(midterm_101)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    79.0    85.0    88.0    87.3    89.0    95.0
```

We can create a boxplot of midterm scores simply by applying the function boxplot() to the vector *midterm_101* :

```
boxplot(midterm_101,
        main = "Boxplot of Dis 101 Midterm Scores",
        ylab = "scores")
text(0.66, 85, "First Quartile (25th percentile)")
text(0.66, 88, "Median (50th percentile)")
text(0.66, 89, "Third Quartile (75th percentile)")
text(0.83, 79, "Minimum")
text(0.83, 95, "Maximum")
text(1, 82,"First Quartile Group", col = "blue")
text(1, 86.5,"Second Quartile Group", col = "blue")
text(1, 88.5,"Third Quartile Group", col = "blue")
text(1, 92,"Fourth Quartile Group", col = "blue")
```



Some definitions for understanding the “Boxplot of Midterm Scores”:

- **The three horizontal lines on the box are called quartiles.** The quartiles divide the data into four “quartile groups,” each containing 25% of the data. In our example, the second quartile group contains the higher 25% of all scores below the median.
- **The dotted lines are called the whiskers,** and therefore boxplots are also known as whisker plots. In our example, the upper whisker extends from the third quartile to the maximum, so it contains the highest 25% of the scores. The lower whisker extends from the first quartile to the minimum, so it contains the lowest 25% of the scores.
- **The second quartile is the median,** which is the thick line inside the box. In boxplots, the median serves as the center of the distribution because it seperates the lower half of the data from the upper half.
- **The vertical length of the box represents the IQR.** The IQR is the interquartile range, which is the difference between the first and the third quartile. That is, the box contains the middle 50% of the scores.

What Do Boxplots Tell You?

1. Variation Within Groups

From a boxplot, we can roughly see how a single variable is distributed. In particular, we look at spread and outliers to observe the variation "within" groups.

- **Spread**
  - **Our original example: The Boxplot of Dis 101 Midterm Scores**

    Boxplots tell us about the "spread" of the data distribution. That is, how far the values are spread out. In our previous example, we can identify from the boxplot that all scores range from 79 to 95, and the middle 50% of scores range from 85 to 89. Furthermore, the third quartile group is the narrowest from the boxplot, so we can infer that the score data is quite condensed between the median (88) and the third quartile (89). We can also see this fact from the frequency table of our data:

    ```
    table(midterm_101)
    ```

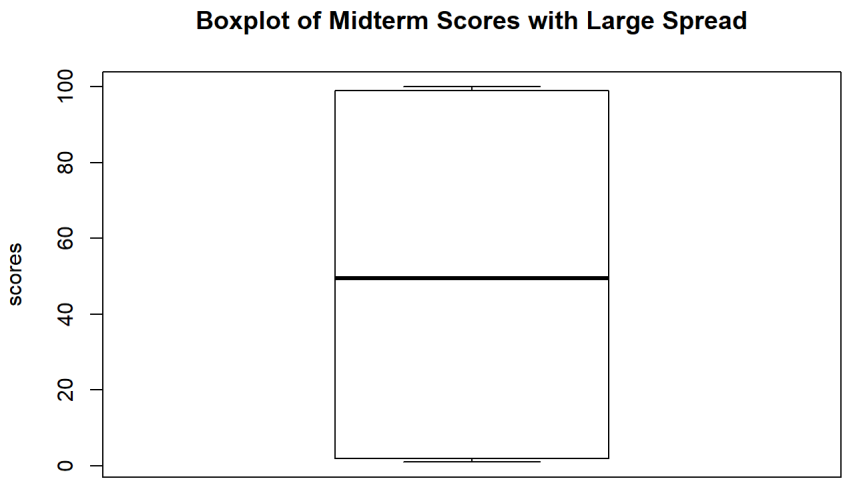
    ```
    ## midterm_101
    ## 79 80 81 82 84 85 86 87 88 89 90 92 93 95
    ##  1  1  1  3  1  2  2  1  5  7  1  2  1  2
    ```

    The table shows that 5 out of 30 students scored 88, and 7 scored 89, meaning that a lot of students scored within this small range, which is consistent with our observation from the boxplot. We can also see from the boxplot that the first and fourth quartile groups are much larger than the second and third quartile group. This indicates that the score data is more condensed around the median, and become sparser as it gets closer to the endpoints.

  - **Data with Large Spread**

    Now consider the case when the boxplot shows large spread:

    ```
    large_spread_score <- c(1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 49, 50, 99,
                    99, 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, 99, 100)
    boxplot(large_spread_score,
    main = "Boxplot of Midterm Scores with Large Spread",
    ylab = "scores"
    )
    ```

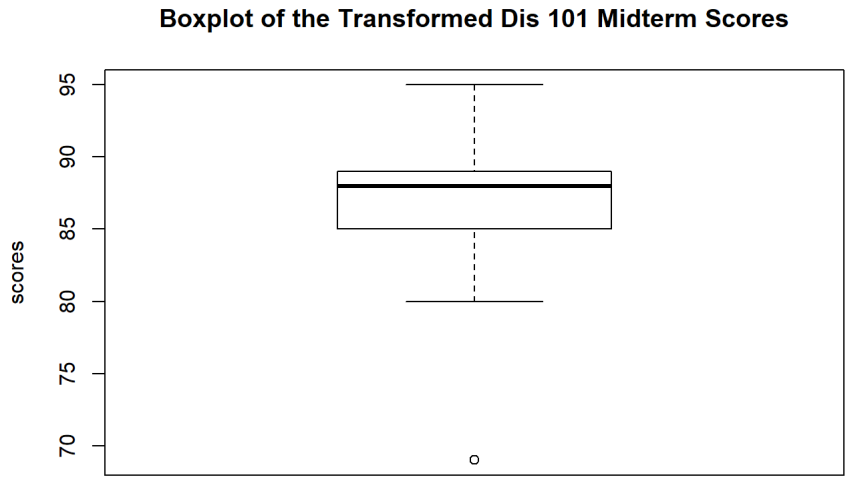
    

    In the above boxplot, the middle 50% of the scores (i.e. the box) is widely spread out. Therefore, knowing that the center (median) is 50 does not tell us much about the data. That is, the center of the distribution is not representative in this case.

- **Outliers**

  When we examine the variation within groups, it is also important to identify the "outliers." An outlier is a value that is too extreme to be considered informative. In R boxplots, the default outlier is any value that is either 1.5 IQR below the first quartile or 1.5 IQR above the third quartile. To create a boxplot with outliers, let's change the lowest midterm score in Discussion 101 from 79 to 69, and make a boxplot with the transformed data:

  ```
  midterm_101_outlier <- c(69, 80, 81, 82, 82, 82, 84, 85, 85, 86, 86, 87, 88, 88, 88,
                      88, 88, 89, 89, 89, 89, 89, 89, 89, 90, 92, 92, 93, 95, 95)
  boxplot(midterm_101_outlier,
      main = "Boxplot of the Transformed Dis 101 Midterm Scores",
      ylab = "scores")
  ```



The empty dot represents an outlier, which corresponds to score 69. This suggests that 69 is much lower than other scores, so it does not tell us much about the rest of the data. In fact, it might even be misleading. Therefore, it's essential to ignore the outlier when we are performing statistical analysis or predictions on the score data. For example, the minimum score shown in the boxplot is the minimum of the score data EXCLUDING the outlier.

2. Variation Between Groups

From a graph containing multiple boxplots, we can compare the distributions of multiple variables. That is, we can observe the variation "between" groups.

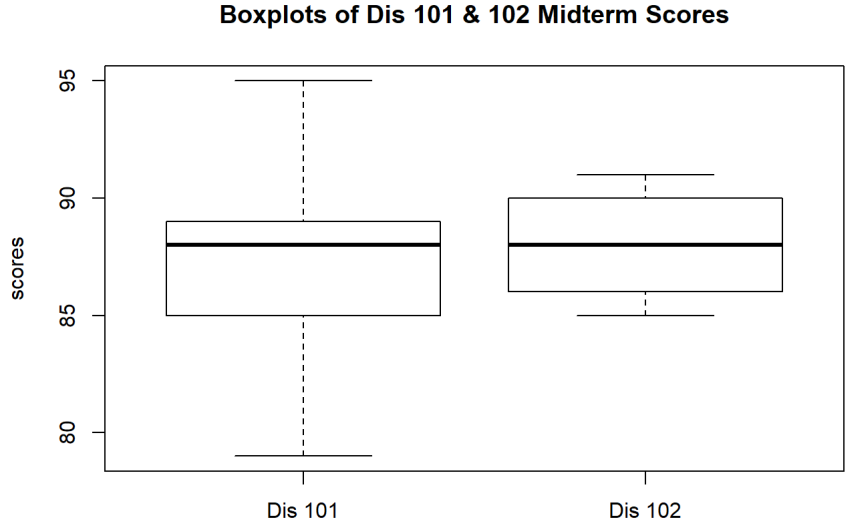
- **Same Center, Different Spreads**

  Consider another set of data from Discussion 102 of Stat133. Suppose there are also 30 students in Discussion 102, and their midterm scores form the vector *midterm_102*:

```
midterm_102 <- c(85, 85, 85, 85, 85, 86, 86, 86, 86, 87, 87, 87, 87, 87, 88,
             88, 89, 89, 89, 89, 89, 90, 90, 90, 90, 91, 91, 91, 91, 91)
```

  Let's compare the boxplots for Discussion 101 and 102:

```
boxplot(midterm_101, midterm_102,
    main = "Boxplots of Dis 101 & 102 Midterm Scores",
    ylab = "scores",
    names = c("Dis 101", "Dis 102"))
```



From the two boxplots, we see that both variables have their center (median) at 88. However, they have different spreads. The Discussion 101 data has a larger range, meaning that the values are (on average) further from the center. That is, students in Discussion 101 scored quite differently on the midterm. The Discussion 102 data has a smaller range and narrower quartile groups, meaning that the values are more condensed about the center. That is, students in Discussion 102 had similar performance on the midterm.

Also, observe that for the Discussion 102 data, the median is roughly in the middle of the box. This type of data with symmetric boxplot is

said to have the “symmetric distribution” because the data is evenly spread out about the median. On the other hand, the Discussion 101 data has the “skewed distribution,” which means that the data is either more condensed on smaller values or larger values. In this case, the median is closer to the third quartile, so the students in Discussion 101 have their scores condensed on higher values.

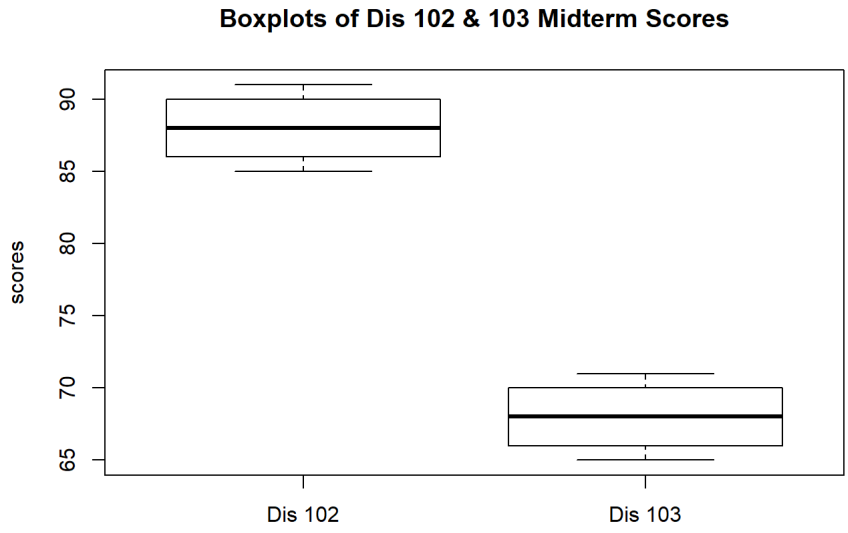
- **Same Spread, Different Centers**

  How about variables with different centers? Suppose there is another section, Discussion 103, with 30 students. Their midterm grades form the vector *midterm_103*:

```
midterm_103 <- c(65, 65, 65, 65, 65, 66, 66, 66, 66, 67, 67, 67, 67, 67, 68,
                 68, 69, 69, 69, 69, 69, 70, 70, 70, 70, 71, 71, 71, 71, 71)
```

  Let’s compare the boxplots for Discussion 102 and 103:

```
boxplot(midterm_102, midterm_103,
    main = "Boxplots of Dis 102 & 103 Midterm Scores",
    ylab = "scores",
    names = c("Dis 102", "Dis 103"))
```



  From the two boxplots, we see that both variables have the same spread. In fact, they both have symmetric distributions. However, they have different centers (medians). The centers for the Discussion 102 and 103 data are at 88 and 68, respectively. Because of this difference, we can see that the boxplot of the Discussion 102 data is much higher than the boxplot of the Discussion 103 data. That is, students in Discussion 102 had higher overall performance on the midterm than students in Discussion 103.

Lastly…

Boxplots may not look familiar, especially for non-statistics majors. Yet, I have demonstrated above how useful they can be. Next time when you see a boxplot, just remember – it’s all about VARIATION!

References

- Box Plot: Display of Distribution: http://www.physics.csbsju.edu/stats/box2.html
- Understanding and interpreting box plots: https://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots
- Interpret the key results for Boxplot: http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/graphs/boxplot/interpret-the-results/key-results/
- WHAT A BOXPLOT CAN TELL YOU ABOUT A STATISTICAL DATA SET: http://www.dummies.com/education/math/statistics/what-a-boxplot-can-tell-you-about-a-statistical-data-set/
- How to Read and Use a Box-and-Whisker Plot: https://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/
- Outlier Treatment: http://r-statistics.co/Outlier-Treatment-With-R.html
- R Box Plot: https://www.programiz.com/r-programming/box-plot