Question 3: Comparison of Two Sub-Populations

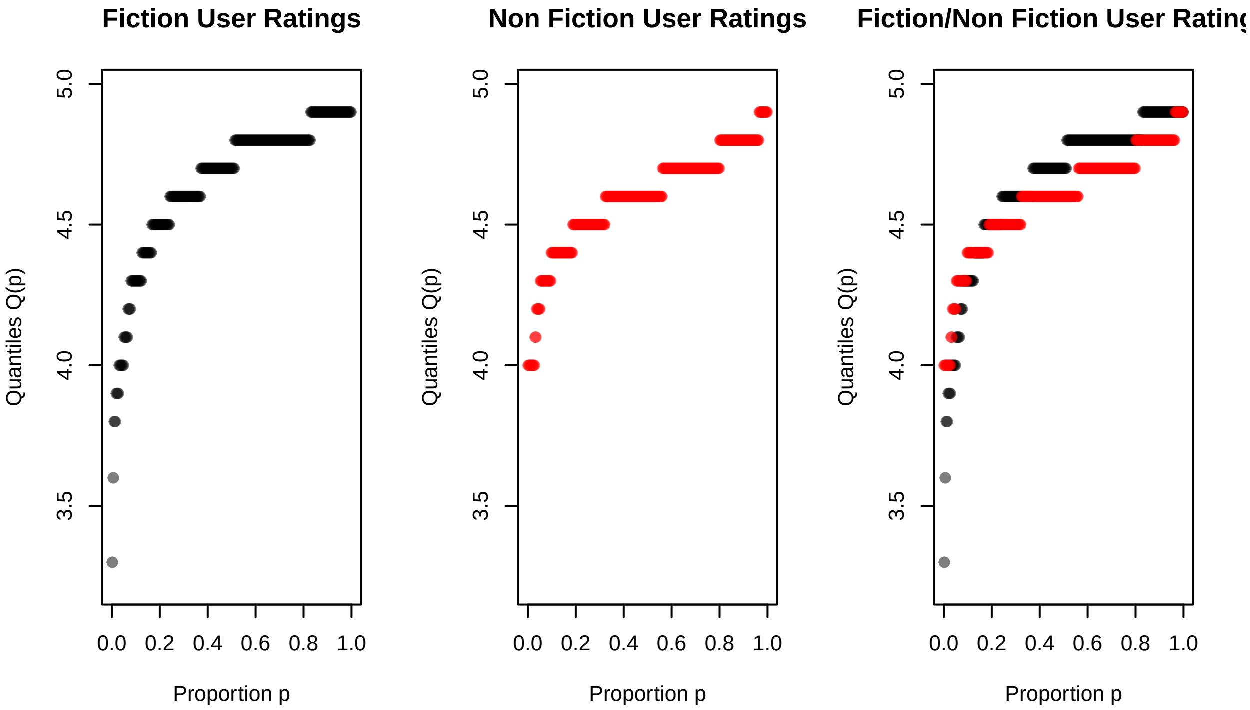
Description of the context and the Two Sub-Populations: The dataset used for this question was found on Kaggle and is titled “Amazon Top 50 Bestselling Books 2009 - 2019”. This dataset includes information on a population, that being the top 50 bestselling books in every year from 2009 to 2019 on the e-commerce site Amazon. Therefore, in total, there are 550 books in this population. This data has been categorized into two genres - fiction and non-fiction - using Goodreads which is a subsidiary of Amazon that has a database of books. The two sub-populations being compared in this question will be Fiction and Non-Fiction books. The dataset itself contains information about the title, author, Amazon user rating, number of reviews on Amazon, year of ranking, price, and genre (fiction or non-fiction).The two attributes used to compare these sub-populations are the population mean and the standard deviation of the user ratings.

Population 1 is Fiction books and Population 2 is Non Fiction. We can numerically compare the user ratings for both sub-populations:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.300   4.600   4.700   4.648   4.800   4.900
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.000   4.500   4.600   4.595   4.700   4.900
```

The first sub-population has a mean of 4.648 and the second population has a similar yet lower mean of 4.595. The first sub-population seems to have a larger range (4.9 - 3.3 = 1.6) than the second sub-population (4.9 - 4 - 0.9). We can also compare these sub-populations graphically using a quantile plot:



Comparing these figures, it seems as though at Q(1/2) or the median, both populations have a similar value. However, Population 1 of Fiction books have a slightly higher value. It also seems as if the 1st quantile

Q(1/4) and the third quantile Q(3/4) occur at similar values at well with Population 1 having slightly higher values. Overall, the general shape of the quantile plots is very similar.

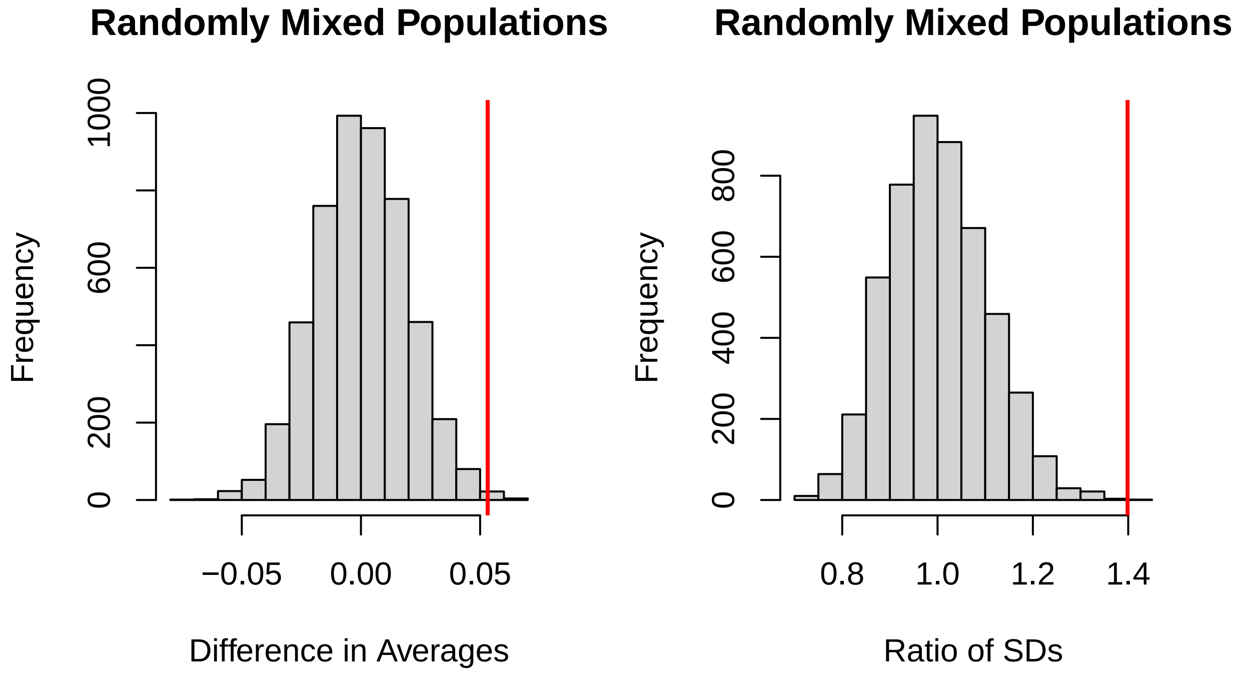
From the histograms of the original two sub-populations, which are not shown to due to spacing limits, it is seen that both of these populations are right-skewed. However, population 1 has much lower tail values towards the left than population 2.

We can also compare these sub-populations by finding the difference of their means. This was calculated to be 0.053. They can also be compared by the ratio of their standard deviations (SDs) which was calculated to have a value of 1.399. To discern whether these values are large or small, we can randomly mix these sub-populations and calculate them again.

These are the differences of means and ratio of SDs of the randomly mixed sub-populations.The difference of means had a value of 0.035 and the ratio of SDs had a value of 0.909.

```
## [1] 0.035 0.909
```

From the values for the original populations, these values do not seem very different. The difference in mean has only decreased by 0.018 units and the ratio of SDs has decreased by 0.49 units. To further investigate, we can shuffle the populations in 5000 different rearrangements and calculate the difference of means and ratio of SDs for each such rearrangement.



Conclusion: The red line on the left plot represents the difference between the average user rating in each of the original sub-populations. The red line on the right plot represents the ratio of the standard deviations of the user ratings in each of the original sub-populations. From the graph on the left and the right, it seems as though these value are extreme relative to the randomly mixed differences and ratios. This reflects that the sub-population features may not be similar as was previously found when just taking one rearrangement. This is why it is critical to take many different rearrangements. Thus, it can be concluded that the sub-populations observed cannot be said to be similar in terms of averages and standard deviations as swapping the units dramatically changes the features of the resulting sub-populations.