STAT7026 Assignment 1 Report

Rui Qiu

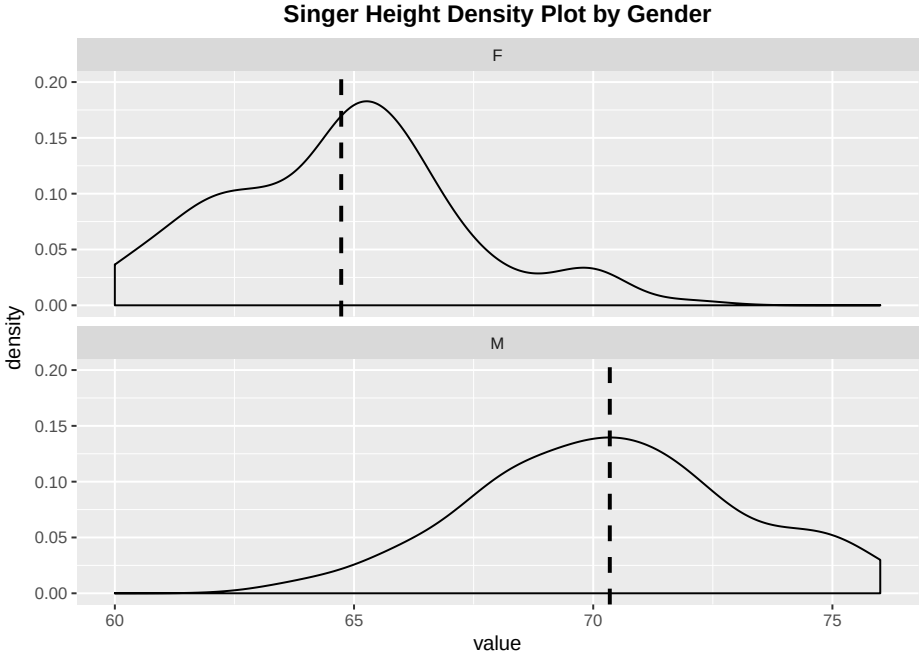
2017-08-18

Background

The heights of singers data from New York Choral Society (1979) contains components `soprano.1`, `soprano.2`, `alto.1`, `alto.2`, `tenor.1`, `tenor.2`, `bass.1` and `bass.2` which are in decreasing pitch order. The first four are female voices and the last four male. We are interested in the comparison of the height distributions, and what factors appear to be important in describing height. We transformed the ragged list into a 3-column dataframe. Those 3 columns represent the height (in inches) of the singer, the voice/pitch of the singer and the gender of the singer.

Visualization by Gender

There are 3 variables in this question, *height, voice(pitch) and gender*. As *height* is the response, we would like to start from *gender*. Conventionally, *gender* is always a good classifier though sounds a little bit politically incorrect. Indeed, the *gender* should contain more information about *height* as the classification by pitch contains some hidden information that all four groups of female voices are higher than those of male in pitch.

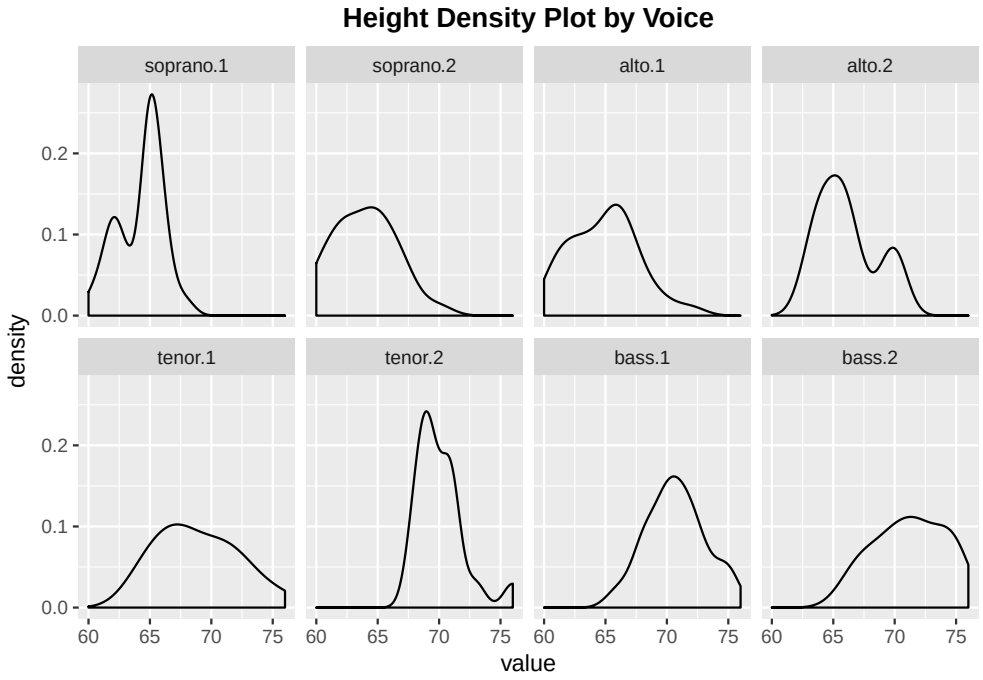


To start visualizing the distribution by gender, we choose **density plot** over histogram since the dimensions of two groups are not equal. The female groups has 127 entries while the male only has 101. Recall the fact that the group sizes of sample actually affect the shape of histogram. So if we want to get an intuitive impression about height distributions of two groups, density plots are our preferable choice.

The **mean** of each group is plotted as dashed line in the plot above. Since the scale of two subplots are identical, we can compare them directly. In fact, we are convinced that **male singers on average is taller (mean is above 70) than female singers (mean is below 65)**, because the male height distribution is generally right-skewed while the female height distribution is generally left-skewed.

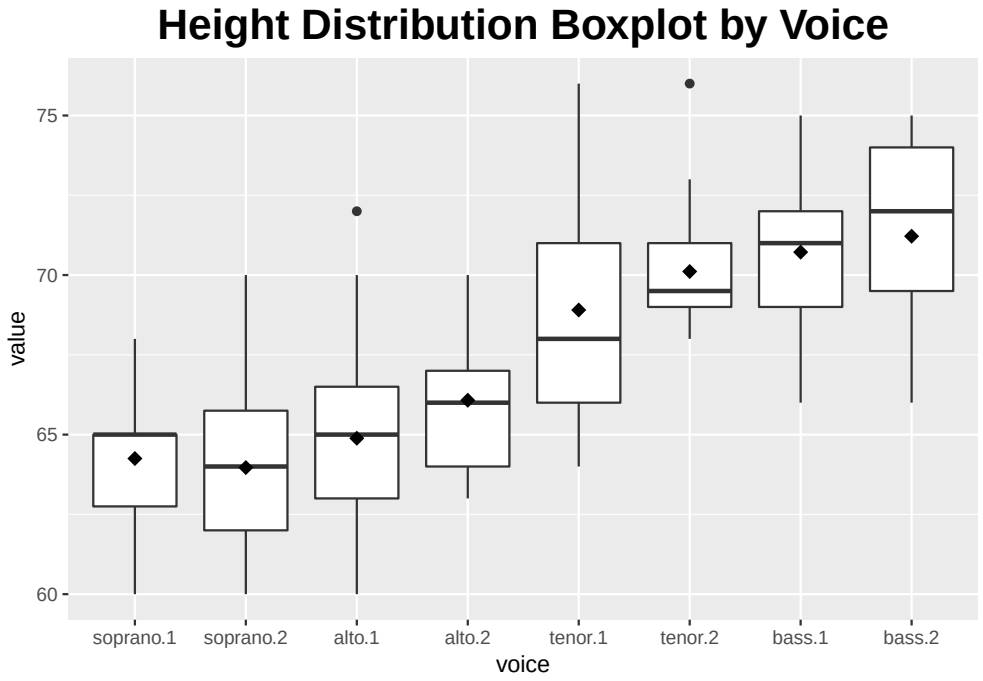
Visualization by Voice

Now we switch to the other variable *voice*. For the same reason, we choose density plot rather than histogram. (In fact, the problem is more severe if we plot the histogram, because group like `tenor.2` is only half the size of group `soprano.1`, then the histograms of the former would be very flat and really hard to identify.) This time, we plot the 8 groups of voices into a 2-by-4 grid so that the first row only contains female singers.



By checking the density plots for each voice group, we may conclude that each voice group follows a **unimodal distribution**. (Although `soprano.1` and `alto.2` could be referred as bimodal distributions, we would rather strictly follow the definition of unimodality in this case.)

Why are we so eager to confirm that they are unimodal? Because it is the precondition for us to construct **boxplot** to compare all eight groups in the same plot.



As we can see, the boxplot helps us examine all 8 groups of data at the same time. The **means** are marked as black squares, and the **medians** are the bold lines in each box. We also rearrange the categorical variable *voice* by decreasing pitch order on x-axis.

Now if we move along the x-axis, which means the voice gets lower, then the corresponding distribution indicates a larger height. Let's rephrase this general trend again, that is, **the height of a singer and the voice of his/hers are negatively correlated**.

Notice that this discovery is not 100% accurate since the group `soprano.1` has both larger median and mean than `soprano.2`. Hence, a refined addition to the discovery above states as: **the pitch affects height more within male group that it does within female group**.

Further Investigation

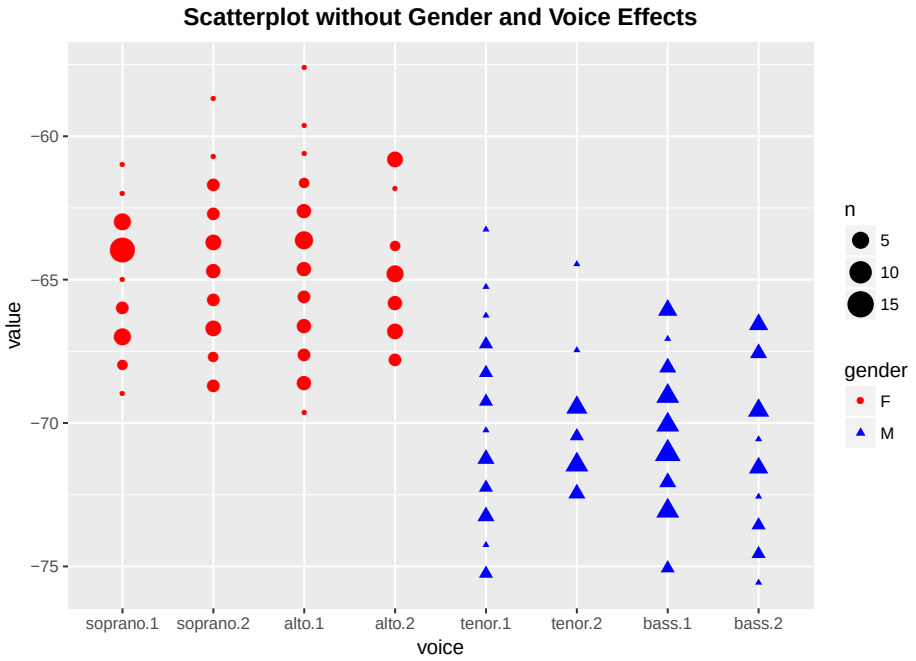
But one would wonder, is that all? Probably not!

So far we've confirmed the roles *gender* and *voice* play in *height*, but we want to see if there is anything extra hidden behind the curtain. We can achieve such goal by constructing a proof by contradiction.

Suppose there is only *gender* effect and *voice* effect in describing the heights, and they are independent with each other. If we eliminate both effects, ideally, we would have some stochastic components left with no pattern. An informal expression is shown below:

$$\text{height} = \text{intercept} + \text{gender effect} + \text{voice effect} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2)$$

What we do next is to subtract the group means (gender means and voice means) from the original height values. The results might seem odd as they are negative. This is because we subtract one more intercept from the height. Since all values are moved down a distance of intercept along y-axis, it doesn't prevent us from checking the "random noises".



Surprisingly, we find out that the scatterplot now displays a seemingly obvious pattern: the female values are generally larger than male values. So there is indeed some extra elements affect the height. Moreover, the *gender* seems involved in this element. Therefore, the interactions between *gender* and *voice* is highly suspicous, but we cannot quantify it with this graph. However, a further stuy with two-way ANOVA might help us figure this out. This process also involves model selection and other tedious actions, so we are not going to expand this topic here.

Conclusions

According to previous visualizations, we conclude that the followings are important in describing *height*:

1. The gender of a singer is correlated with height. Male singers tend to be taller than female singers.
2. The height of a singer is negatively correlated with the pitch of his/hers, i.e., a singer with lower voice tends to be taller.
3. Some hidden elements also affects the height, which seems related with *gender* again.

One **intuitive explanation** is that we can consider human as a piccolo, the shorter air column it has, the frequency of vibration is higher, so that the pitch is higher. This is why, in most cases, male singers with `bass` voice are always taller than female `sopranos`.

Since the two variables are categorical, we can always use logistic regression to contruct a linear model in details. We should also notice that a multiplicative model with interaction term is appropriate for this data.

References

- R for Data Science by Garrett Grolemund and Hadley Wickham, http://r4ds.had.co.nz/