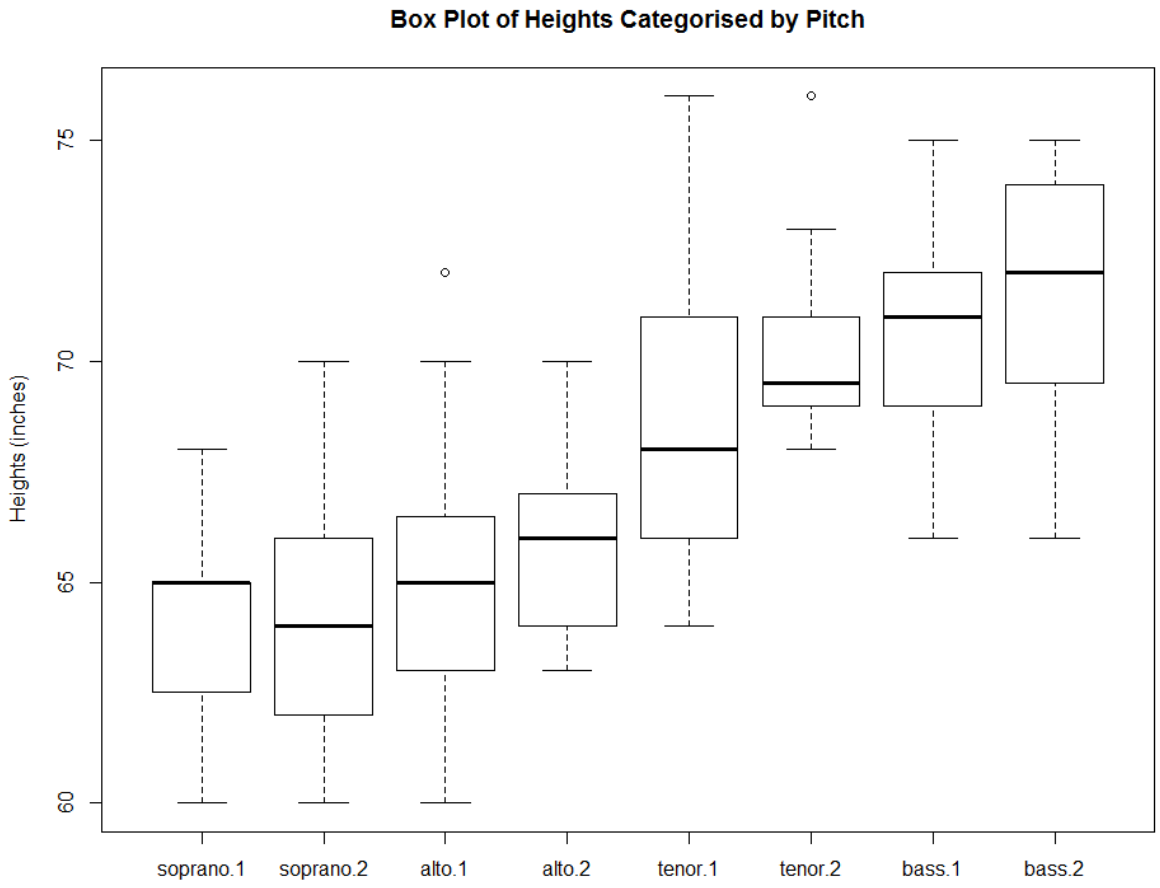
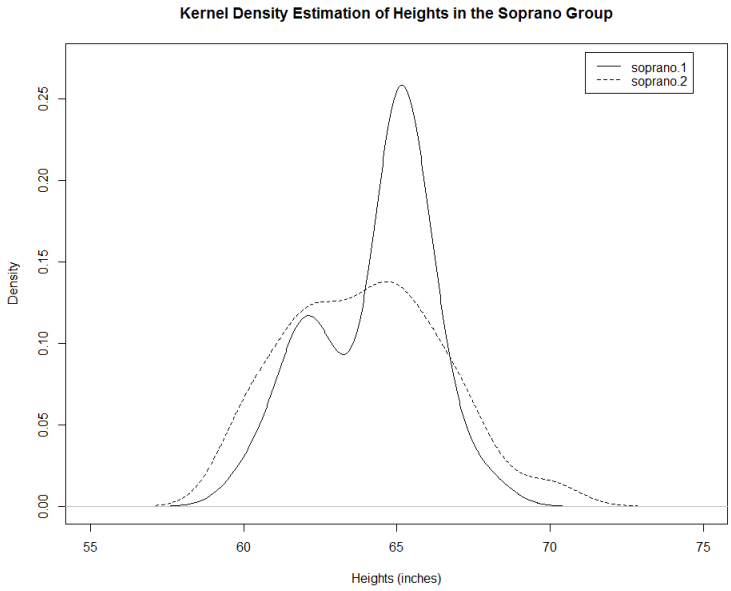
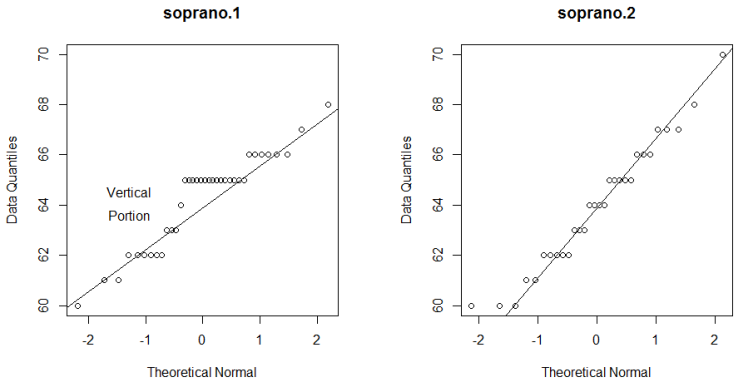


- The box plot above is used to **compare the height distributions of singers** in the New York Choral Society in 1979 **based on pitch listed in descending order**, with the highest pitch being soprano.1 and the lowest being bass.2.
- We can observe some **features of the singers' data** through this box plot:
  1. There is an **outlier** in the **alto.1** group and in the **tenor.2** group, showing that there is an **exceptionally tall singer** in each of both groups.
  2. The **interquartile range** of height distributions in different pitch groups **does not follow a fixed pattern**, the **variability of singers' heights are random** in each group, with **tenor.1** having the **greatest variability in height** and **soprano.1** having the **least variability in height**.
  3. A very important observation is that the **median of singers' heights** generally **increases as pitch decreases**, with the **exception** of the group of singers in soprano.1.
     - We can see that different from the rest, the **median of singers' heights in soprano.1 is higher than that in soprano.2**. The median of the soprano.1 group also happens to be equal to its 3rd quartile.
     - Hence, we are going to take a closer look at the **kernel density estimation** and the **quantile-quantile plots** for the **soprano.1** and **soprano.2** data in the next page.

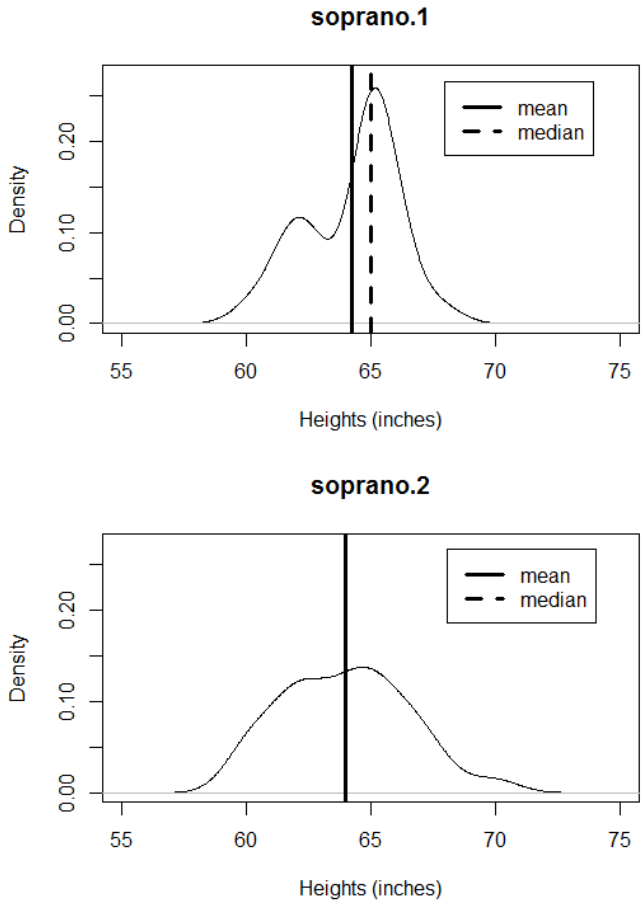


Quantile-Quantile Plots of Heights in the Soprano Group



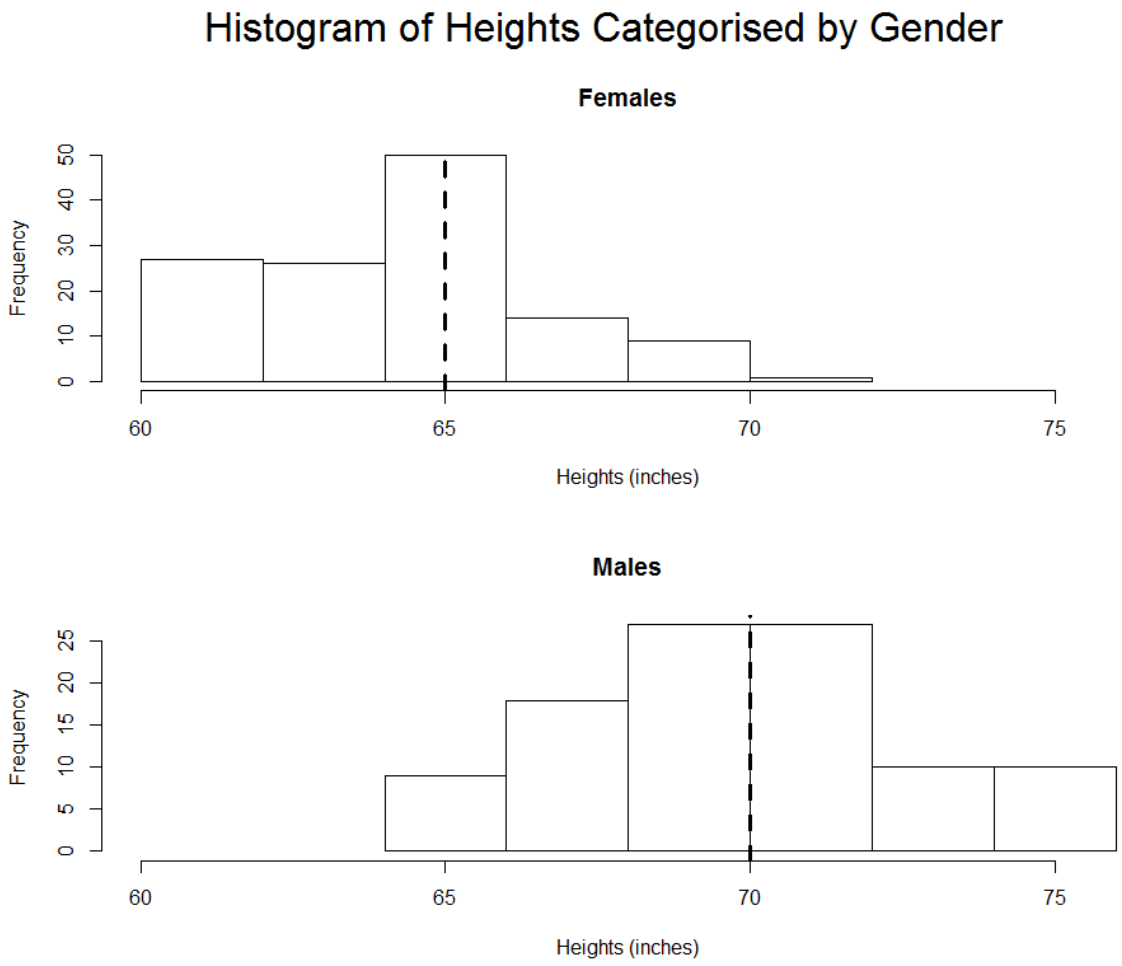
- The first diagram at the top of this page shows the kernel density estimation for the soprano.1 and soprano.2 height data.
- The second diagram shows the respective quantile-quantile plots.
- From the kernel density estimation above, we can see that the distribution of the **soprano.1** height data looks like a **bimodal** distribution.
- Looking at the quantile-quantile plot for soprano.1, we can see a **vertical portion** (labelled with the words "Vertical Portion"). This vertical shape might not look very obvious because the singers' heights are rounded (data aggregates at integers), and so the data appears as flat steps. However, that **vertical portion indicates that the soprano.1 data is indeed bimodal**.
- This suggests that the box plot did not represent the soprano.1 data well, because a key assumption when representing data with a box plot is that the data set is unimodal.
- However, in my opinion, a box plot is still the best way to compare the height distributions between 8 different pitch groups because comparing 8 histograms or 8 density plots at the same time will not be very effective, as comparing height distributions and observing patterns will become more difficult and complicated when looking at 8 different plots.
- We can also observe a few more features of the data from the quantile-quantile plots. We can see that the **soprano.2 data follows the straight line quite closely**, indicating that it is **normally distributed**. Again, this might not be very obvious as the singers' heights are rounded. Also, the **slope of the soprano.1** data is **less steep** than that of the **soprano.2** data, indicating that the **soprano.1 heights are not as spread out as the soprano.2 heights**.

Kernel Density Estimation of Heights in the Soprano Group



- The diagram above shows the kernel density estimation for the soprano.1 and soprano.2 height data separately. The **means and medians** of the heights are represented by **solid vertical lines** and **dotted vertical lines** respectively.
- From the kernel density estimation above, we can see that the **mean** height of **soprano.2** is **exactly the same as its median**. This is a general characteristic of data that is **symmetrical**, which agrees with our earlier observation of the quantile-quantile plot for soprano.2, that the soprano.2 heights are normally distributed.
- For **soprano.1**, we can see that the **mean** height is **less than its median**.
- The **difference between the means** of soprano.1 and soprano.2 is actually **smaller or less significant than the difference between their medians** as can be observed in the density plot above.
- Since the **mean** is a **better measure** of central tendency to use to compare the height distributions of soprano.1 and soprano.2 (because soprano.1 data is bimodal), thus, our initial idea that the centre of soprano.1 is significantly higher than soprano.2 (from the box plot observation that the median of soprano.1 is greater than that of soprano.2) is actually not true. Their centres are in fact quite close to each other, as observed from the means of both data sets in the density plot above.
- Although there are a few exceptions or special cases (such as some outliers observed and the centres of soprano.1 and soprano.2 being quite close to each other), from the box plot, we can **conclude that in general, as pitch decreases, the heights of singers increase**.
- Thus, we can **conclude that pitch appears to be an important factor in describing height**.

- We also know that the **first four groups** (soprano.1, soprano.2, alto.1 and alto.2) consist of **females** and the **last four groups** (tenor.1, tenor.2, bass.1 and bass.2) consist of **males**.
- Hence, we can now analyse **another potential factor, gender**, which may also appear to be important in describing height.



- The histograms above are used to **compare the height distributions** of the singers **based on gender**.
- We can see that the **height distribution** of **female** singers is **skewed to the right**.
- The **dotted vertical lines** on both histograms represent the **medians** of the heights of **female** and **male** singers. Since a skewed distribution will have a mean that is distorted by the long tail on one side, we compare the medians of both gender groups.
- Clearly, the **histogram for males** sits to the **right** of the **histogram for females** and the **median height of male singers** is **greater** than that of **female singers**.
- Hence, we can say that **in general, males are taller than females**.
- Thus, we can also **conclude that gender appears to be an important factor in describing height**.
- In conclusion, both pitch and gender are important factors in describing height.