Discussion Solution: Biometry student height, age and piercings

A note on write-ups: In most of my ‘solutions’ I will provide much more detail than you need to do, and will include graphs and tables of statistics which you do not need to show. For this week, and perhaps occasional others, I first give an example of the sort of concise verbal write-up you should submit, and then give details.

Concise write-up

To describe the distributions I used histograms, boxplots and basic descriptive statistics. Normal quantile-quantile plots were used for any distributions which appear to possibly be approximately Normal (i.e. are symmetric and bell shaped). Five-number summaries were used as descriptive statistics unless the distribution was roughly Normal, in which case mean and standard deviation were used.

The distribution of ages is regular, and strongly skewed with a long upper tail and almost no lower tail. The median is 26 years. Most values vary only slightly, between the minimum of 21 and the upper quartile of 31. There is one clear outlier at 64 and seven other moderate outliers (> 40), but all the values are plausible and indeed are consistent with the longer upper tail.

There are four values of height which are highly implausible (< 68 cm) or impossible (5.6 cm); these cannot be corrected with certainty and so should be excluded. The distribution of the remaining values is regular, and slightly skewed with a longer lower tail. The median is 170 cm. The quartiles are fairly symmetric around the median (163 and 177), but the long lower tail and slightly compressed upper tail produce clear curvature in the NQQ plot.

The distribution of piercings is irregular, with most students having 0 or 2 piercings, but very few having 1 piercing. This presumably arose from the bilateral symmetry of the human body and a tendency to have matching piercings, particularly one in each ear. The smaller number of students with > 2 piercings produce a long upper tail, with no values that appear as outliers in the histogram (though several are flagged as extreme by a boxplot).

To compare distributions between genders, among colleges, or classified by both gender and college, I used primarily boxplots.

The distributions of ages are quite similar between the genders, with the principal differences being in the upper tails. Students in CTAHR and ‘other’ are somewhat older — shown mostly in the medians and minima — than students in NatSci and SOEST. Separating ages by both gender and college showed nothing new.

The main differences in distributions of height is that, as expected, males are somewhat taller than females (medians = 179 and 164). The distribution also is more skewed for males, with a short upper tail and fairly narrow range between the quartiles but a long lower tail, including two surprising short males; the distribution for females was broader as measured by the IQR but had short tails on both ends. There is little difference in distributions of height across the colleges, though values were somewhat less variable
for CTAHR students. Separating heights by both gender and college showed that CTAHR males tend to be slightly shorter than the other males, but otherwise the differences were only those already seen between the genders.

As expected, the distributions of number of **piercings** differs greatly between the genders, with more than 75% of males having none and more than 75% of females having at least two. Single piercings are rare in both genders, so the bimodal distribution with peaks at 0 and 2 appears in both. Far more females had more than two piercings, but still nearly half had exactly two. The distributions of piercings showed no major differences among colleges. When broken down by both gender and college, fewer males had any piercings in NatSci and SOEST than CTAHR and ‘other,’ while females in NatSci and SOEST had more variability than females in CTAHR and ‘other.’

**Details**

1. Describe the distributions of the three quantitative variables.

To describe distributions I use histograms, boxplots, and basic descriptive statistics. Normal quantile-quantile plots are useful for any which appear to possibly be approximately Normal (i.e. are symmetric and bell shaped).

**Age:**

- The age distribution has a single primary peak at about 25 years. There appears to be a slight secondary peak of 35 and 36 year old students, though this is at least partly due to a paucity of 33 and 34 year olds; this second peak could represent a real heterogeneity in the students who take the class: some have returned to school after some years working, while most have come to graduate school soon after graduating from college, if not immediately. This second peak, though, is small enough that I think it is reasonable to consider the data as having a single distribution and to use measures of center to describe it.
- The distribution is strongly skewed, with most observations between 21 and about 30 and a long right tail of a moderate number of students in their 30s and a few in their 40s to 60s. Because of this skew I feel the five-number summary is better for summarizing the distribution than are the mean and standard deviation.
- The median age is 26; because of the skew, the mean is somewhat larger (28.4).
• The main part of the distribution is fairly narrow, with the middle half extending only from 24 to 31 (Q1 and Q3); because of the long upper tail the total distribution is much wider (21 to 64).

• The largest value (oldest student) does lie apart from the rest of the distribution in the histogram, and the largest seven are classified by the boxplot as possible outliers, but none of these values is implausible and indeed none seems inconsistent with the pattern of the upper tail. Note in particular that it would be nonsensical to regard the seven largest values, including a 42 year old, as outliers but not the eighth largest (41 years old).

Height:

A histogram or boxplot of the heights immediately shows a problem: four observations are implausibly small (67 cm or shorter). There are reasonable explanations for the mistake(s) made in recording these heights, but since there is more than one such possible interpretation, it would be unwise to try to “correct” these values. Fortunately they are few enough, relative to the sample size, that omitting them (rather than converting them in some reasonable way) has little effect.

All following analysis excludes these four implausible/impossible values. The distribution without them is:
• The distribution has a single main peak at around 170 cm. There is a second peak at 157 cm (which actually is the most frequent single value), but as with the age variable, I think it still is reasonable to regard the data as having a single distribution for which measures of center are meaningful.

• Even ignoring the outliers, the distribution is slightly skewed, with the upper tail shorter than the fairly spread-out lower tail. This is evident also in the clear, if slight, curve in the Normal quantile-quantile plot shown to the right.

• The median is 170 cm (with or without the outliers); the mean is slightly smaller than the median both with and without the outliers, but more so with the outliers included, indicating that the outliers rather than the asymmetry of the rest of the distribution have the greater impact on the mean.

• The distribution is fairly broad, with the middle half of the values ranging from 163 cm to 178 cm (i.e. an IQR of 15 cm or nearly 6 inches); excluding the outliers increases the lower quantile somewhat but has no effect on the upper quantile.

• If the outliers are excluded, the next smallest value after the outliers (130 cm) is slightly separated from the others in the histogram and is classified as a possible outlier by the boxplot. It clearly is not an implausible or even highly surprising value, and indeed I don’t think it is inconsistent with the shape of the rest of the lower side of the distribution.

**Piercings:**

• The distribution of numbers of piercings clearly has multiple peaks. There are two non-exclusive interpretations of the pattern. First, it might be reasonable to regard the students as being of two types: those with piercings and those without. In other words, the peak at 0 represents a different subpopulation than the rest of the data, which has a fairly regular distribution. In addition there are more observations with even numbers
of piercings (at least up to 4) than with adjacent odd numbers; this presumably is because most piercings, at least people with only a few piercings, are for earrings and for both ears.

- The distribution is very skewed, with the minimum value, 0, also being the most frequent, and numbers greater than 2 being fairly uncommon.

- The median is 2.

  Note that nearly half the observations (61 of 140) are 0 or 1, so that with only a slight change in the distribution the median would be 1; this in fact was the case in previous years. Because of the even/odd gappiness of the distribution, though, it wouldn’t make sense to say that a median value of 1 represents the “typical” value.

  Note also that despite the general long-right-tail skew in the distribution, because of the rarity of observations with 1 piercing, the median actually is slightly larger than the mean.

- The distribution is fairly narrow, with nearly three-fourths of students having 2 or fewer piercings.

- The boxplot classifies the two observations with 9 piercings and one with 8 as outliers (it is hard to see but there are two dots at 9, slightly offset vertically). On the histogram these points are only mildly exceptional given the long right tail of the distribution. On the other hand, clearly those with more than 4 piercings are in a small minority (14 of 140).

2. Compare these distributions between sexes and among colleges.

For comparing distributions between or among groups, I primarily use boxplots and descriptive statistics; comparing histograms is useful also, if you want to compare the detailed shapes of distributions as well as their centers, spread, and degree of skew.

**Age by sex:**

The distributions of ages for female and male students are very similar. The minima, medians, and upper quartiles are identical, and the female lower quartile is lower than that for males by only 1 year. The only substantial difference is that the male distribution had only one unusually large value (one student more than 60 years old) while there were four females over 45 years old.

**Age by college:**

The distributions of ages over the four college categories are broadly similar in level and shape. Students from CTAHR and “other” (mostly Social Sciences), however, were slightly older and somewhat more variable in age than were students from CNS and SOEST. The minima for the latter two were slightly lower than for the former two, and this difference became progressively larger for Q1, median, and Q3.

The other noteworthy difference is that there were four CNS students markedly older than the others from that college, while there were no major outliers in the other colleges.
Age by sex and college:

Since there was little difference in ages between the sexes or among the colleges, little more is seen when the observations are grouped by both sex and college. The main thing to observe here is that the older CTAHR and “other” students are females, and the same is true for CNS students apart from the single 64 year old male.

<table>
<thead>
<tr>
<th>Variable</th>
<th>college</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Min</th>
<th>Q1</th>
<th>Med</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>CTAHR</td>
<td>26</td>
<td>29.92</td>
<td>6.090</td>
<td>23.00</td>
<td>24.75</td>
<td>28.50</td>
<td>35.00</td>
<td>48.00</td>
</tr>
<tr>
<td></td>
<td>NatSci</td>
<td>69</td>
<td>27.81</td>
<td>7.387</td>
<td>21.00</td>
<td>24.00</td>
<td>26.00</td>
<td>29.50</td>
<td>64.00</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>22</td>
<td>29.86</td>
<td>6.600</td>
<td>22.00</td>
<td>24.75</td>
<td>28.50</td>
<td>33.25</td>
<td>47.00</td>
</tr>
<tr>
<td></td>
<td>SOEST</td>
<td>21</td>
<td>27.00</td>
<td>3.912</td>
<td>21.00</td>
<td>24.50</td>
<td>25.00</td>
<td>30.00</td>
<td>36.00</td>
</tr>
</tbody>
</table>
For comparisons of heights between sexes or among colleges I excluded the four clearly invalid values. As was to be expected, males generally are taller than females: Q1, median, and Q3 all differed by 13 to 16.75 cm. The shapes of the distributions are similar (both fairly symmetric).

The differences in shape are in the extremes. The upper tail of the distribution is shorter for males than females: the difference between the sexes becomes progressively smaller from Q1 through the median and Q3 up to the maximum.

More striking is the presence of two unusually short males: values of 130 and 140 cm that are shorter than any of the plausible values for females (shortest = 141 cm) and much shorter than any other males (next shortest = 156 cm). These two values, which were not clear outliers when the sexes were combined, clearly now are. The boxplot also shows two other males (156 and 157 cm) as outliers, but in a histogram or dotplot they do not look very exceptional.
The IQR for males is somewhat smaller than for females (9.25 vs. 13), while the standard deviations differ by about the same amount but in the opposite direction. This reflects the sensitivity of the standard deviation to the long lower tail of the male distribution, and especially the two low outliers.

**Height by college:**

![Box plot of height by college](image)

The distributions of heights are quite similar among the colleges. The median is slightly smaller, and there is slightly less spread (IQR or total range) for CTAHR students than those of the other colleges.

All four distributions were roughly symmetric, though both SOEST and Natural Sciences had somewhat longer lower than upper tails, and CTAHR was the reverse.

Because height also is related to sex, comparison of heights across the colleges is confounded by the possibility of differences in the sex ratios of the students in the different colleges. This is addressed later by describing height in relation to both college and sex.

<table>
<thead>
<tr>
<th>Variable</th>
<th>college</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Min</th>
<th>Q1</th>
<th>Med</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>height</td>
<td>CTAHR</td>
<td>26</td>
<td>167.69</td>
<td>9.84</td>
<td>150</td>
<td>160.0</td>
<td>168.0</td>
<td>173.5</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td>NatSci</td>
<td>69</td>
<td>171.33</td>
<td>10.66</td>
<td>141</td>
<td>163.0</td>
<td>171.0</td>
<td>180.0</td>
<td>193</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>21</td>
<td>168.57</td>
<td>13.42</td>
<td>140</td>
<td>157.0</td>
<td>171.0</td>
<td>178.0</td>
<td>194</td>
</tr>
<tr>
<td></td>
<td>SOEST</td>
<td>19</td>
<td>166.16</td>
<td>14.09</td>
<td>130</td>
<td>157.0</td>
<td>171.0</td>
<td>177.0</td>
<td>186</td>
</tr>
</tbody>
</table>

**Height by sex and college:**

Separating heights by both sex and college shows mainly the difference between males and females, which is generally consistent across colleges, except that CTAHR males tend to be slightly shorter than males from the other colleges. The differences in skew among colleges noted above are seen here to be largely due to a few exceptional low values, one in each sex in Natural Sciences and one male in each of “other” and SOEST.
Piercings by sex:

As was to be expected, females had more piercings than did males. The distributions are so different, so skewed, and so peculiar (due to the even/odd pattern), that the standard descriptive statistics aren’t very useful. In addition a boxplot is not useful for the males since even Q₃ is 0: there is no box or lower whisker!

- Only 11 males out of 59 (19%) had any piercings, while 84% (68 of 81) of the females had at least one.
- The students with more than 2 piercings also were almost entirely female (36 vs. 2), with 14 females (about 17%) having more piercings than the maximum number for males (one male with 4 piercings) and the female maximum being more than twice the male maximum (9 vs. 4).
- Among females, nearly half those with any piercings had 2 piercings (26 of 59).
Piercings by college:

Comparison of the distributions of piercings across the colleges is complicated by the spiky even/odd shapes of the distributions. Overall there are no major differences evident. The distributions for SOEST and “other” in particular are quite similar. The preponderance of 0s, and the higher frequency of even than odd numbers, both are slightly greater for Natural Sciences than the other colleges. In contrast, CTAHR is the only college in which 0 is not the most frequent value.

These differences among the colleges again can be largely explained by the different sex ratios, as shown later.
Piercings by sex and college

So few males had piercings there is little to say comparing numbers of piercings across colleges for males, the main difference being that fewer Natural Sciences and SOEST males than CTAHR and “other” males had any piercings. Among females, the only notable difference is that the fractions with 2 piercings were much higher for CTAHR and “other” while there was more variability among women in Natural Sciences and SOEST.