Homework Assignment One
Jack L. Vevea

I attempted to produce a "perfect" histogram showing the Raven distribution in my sample. I began by using R's stem() function to suggest an appropriate grouping:

```
> stem(Raven)

  The decimal point is 1 digit(s) to the right of the |

  1 | 3
  1 | 5578
  2 | 1111123334
  2 | 56678888
  3 | 0122334
  3 | 5557888999
  4 | 0112234
  4 | 557
```

With eight categories, that met the 7-15 interval guideline we have discussed in class, so I used that same grouping to produce the histogram:

```
> hist(Raven, breaks=seq(9.5,49.5,5), axes=FALSE)
> axis(side=1, at=seq(12, 47, 5), pos=0)
> axis(side=2, at=seq(0,10,2))
> abline(h=0)
```

The histogram that resulted appears on the following page.

According to either of the common measures of central tendency, the Raven distribution appears have its most typical values somewhere in the low thirties:

```
> mean(Raven)
[1] 30.84
> median(Raven)
[1] 31.5
```

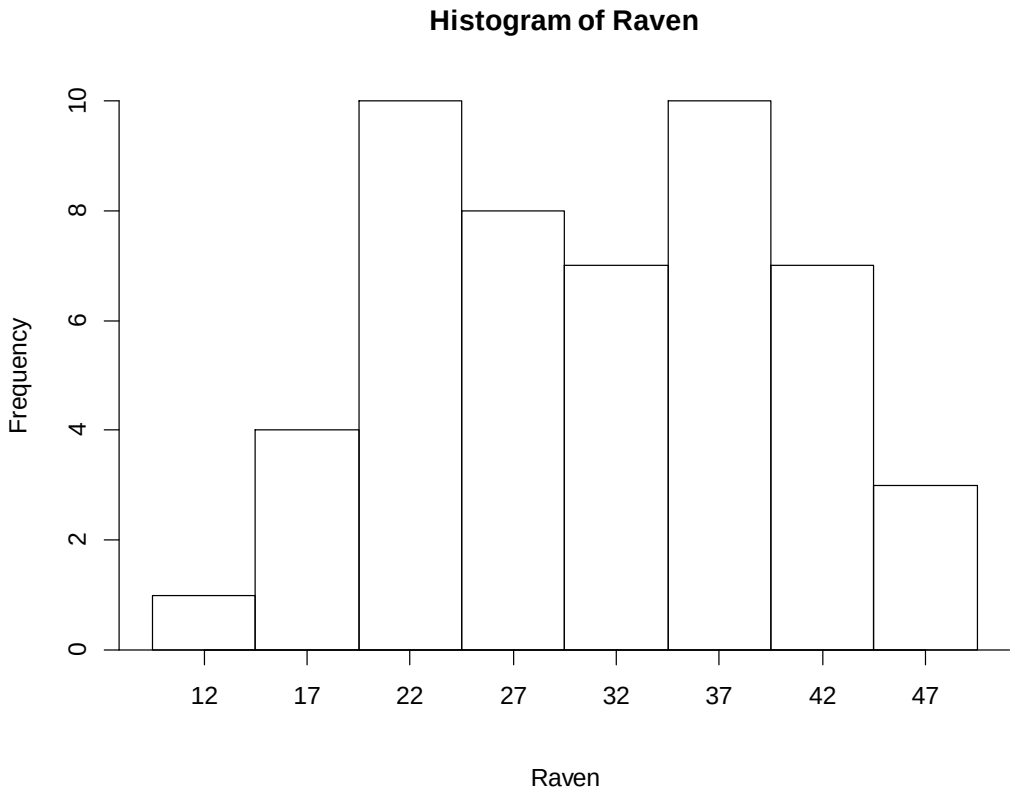
Those values (mean=30.84, median=31.5) are consistent with the graphical evidence in both the stem-and-leaf plot and the histogram. It is clear that the distribution would balance at about 31, and that half of the area in the histogram would appear above or below a value near 31.5.

The distribution of scores is not particularly tightly packed around that center:

```
> sd(Raven)
[1] 9.132494
> iqr(Raven)
[1] 16
```

As there is no compelling reason to consider the median as the most appropriate measure of central tendency here, I would probably focus on the mean, in which case I would note that the standard deviation of 9.1 indicates that the typical observation is about 9 points away from the mean of 30.8. On the other hand, if some concern (e.g. an extreme observation that inflated the mean) caused me to focus on the median, I would note that the central half of the distribution falls within a range of 16 points around the median of 31.5.

The histogram produced in the first step appears to be highly symmetric:



That is consistent with the fact that the mean and the median are about the same. Pearson's scaled comparison of the mean and median results in a value of about -0.2, quite close to zero, which also indicates a symmetric distribution:

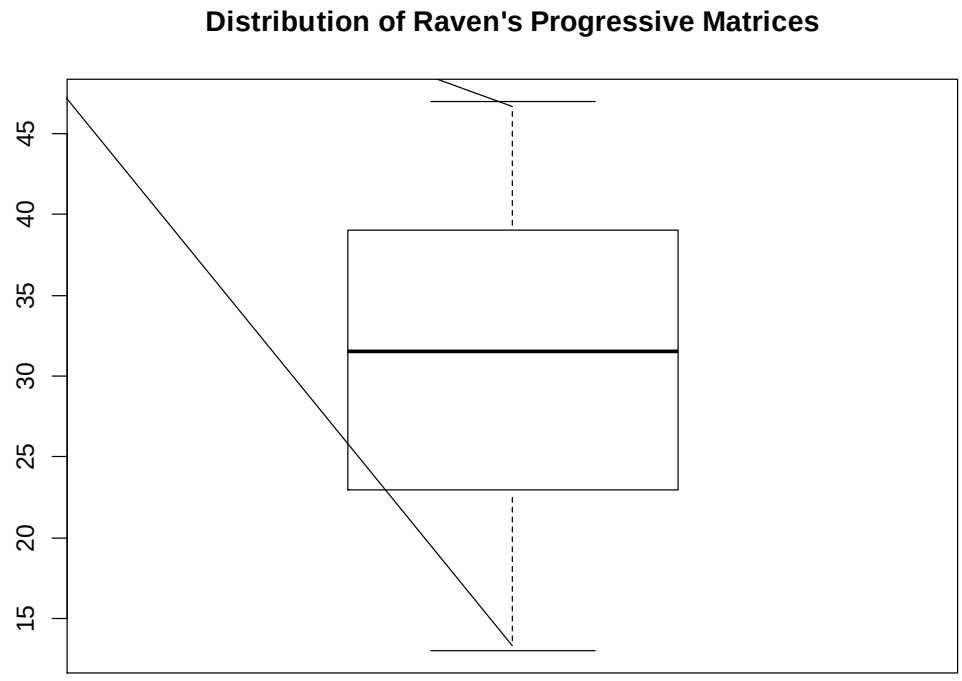
```
> pskew(Raven)
[1] -0.2168082
```

The fact that the third quartile (39) and the first quartile (23) of the distribution are roughly equidistant from the median also indicates that the distribution is symmetric:

```
> quantile(Raven, .75, type=2)
75%
 39
> quantile(Raven, .25, type=2)
25%
 23
> quantile(Raven, .75, type=2)-median(Raven)
75%
7.5
> quantile(Raven, .25, type=2)-median(Raven)
 25%
-8.5
```

The graphical evidence of a box-and-whisker plot gives a nice picture of that symmetry. Note that the central box in the plot has the median line almost exactly at its center, and the whiskers extend almost the same distance above and below the box:

```
> boxplot(Raven, main="Distribution of Raven's Progressive Matrices")
```



One further feature of the histogram is worth noting: There appear to be separate modes in the low 20s and high 30s. However, if even one or two observations were moved to adjacent categories, that appearance would go away. Because this is a relatively small (*N*=50) data set, I do not see this as evidence that the distribution is bimodal.

Another interesting point is that the performance of the children in my sample appears to be below that of the norming sample: the median of 31.5 in my sample is 4.5 points lower than the median of 36 in the norming sample. That's lower by roughly 1/3 of the interquartile range of 13 that described the norming sample. My sample is also about 23% more variable than the norming sample:

```
> 16/13
[1] 1.230769
```

That is not particularly surprising, as these Raven scores were collected in Oakland, CA, which is most likely a community with lower socioeconomic status than would be present in a norming sample.