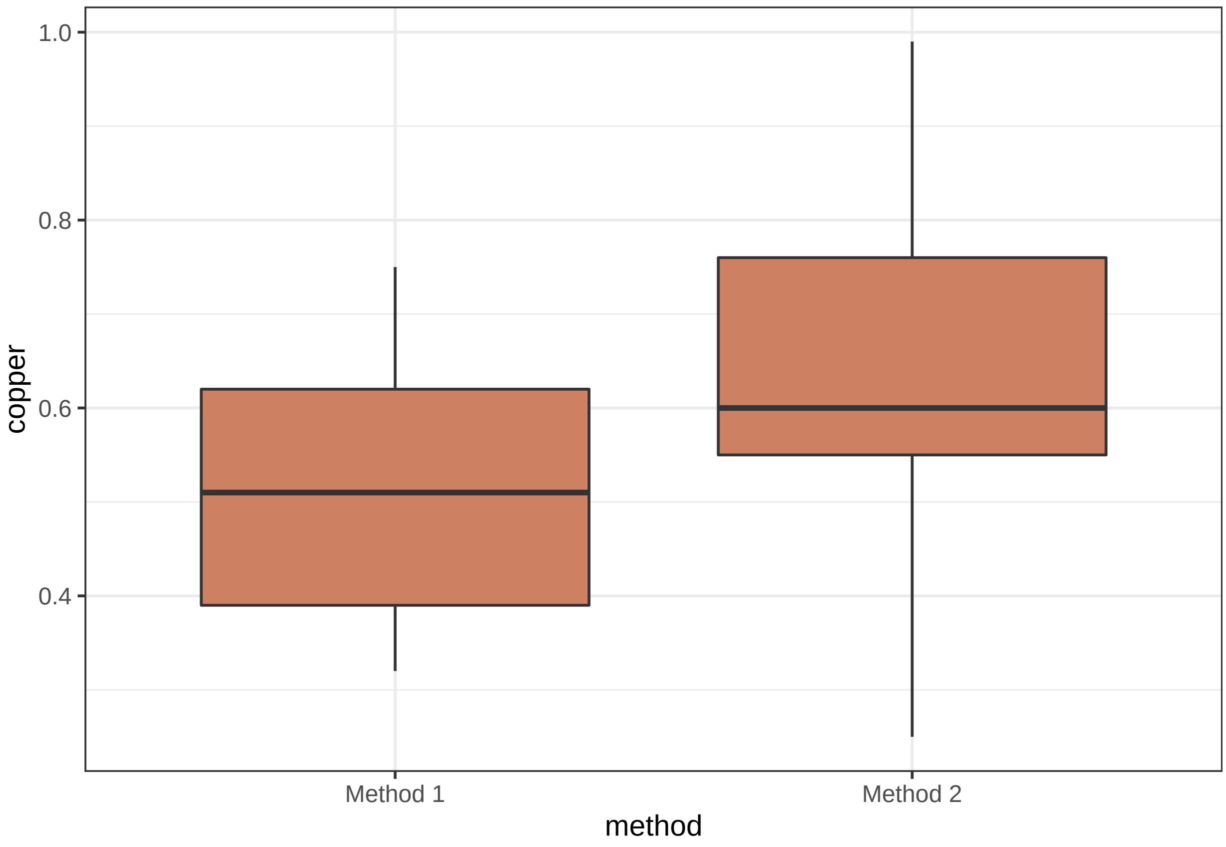
Discussion 2: Numeric and Graphical Summaries

1. Two methods were studied for the recovery of copper. Thirteen runs were performed using each method, and the fraction of protein recovered was recorded for each run. The results are summarized in the plots below:



   a. Compare the shape, center, and spread of the two methods' fraction of copper recovered.

   b. Can you use the above plot to find the mean fraction of copper recovered using Method 1? Explain.

   c. Can you use the above plot to find the mean fraction of copper recovered using Method 2? Explain.

   d. After going over the original data again we find out that there was a data entry error. The maximum of Method 1 was accidentially put in as .75, but should have been 1.75. What would this do to the mean of Method 1? What would it do to the median?

e. If the mean of the data for Method 2 were 0.63, and another observation was added to the Method 2 group with a value of 0.63, would the standard deviation increase, decrease, or stay the same? (HINT: look at the definition of the variance, and think about what happens if you add another term where $x_i = \bar{x}$.)

2. The data included in the lightspeed.csv file shows the results of a set of 5 experiments performed by Albert A. Michelson between 1877 and 1882 to measure the speed of light. These were the first in a series of experiments that eventually disproved the existence of “luminiferous aether”, the hypothesized medium of propagation for light waves. This research set the stage for the theory of special relativity two decades later.

a. Make a histogram using the full dataset. Choose binwidth to be 25. Note that the data is listed in km/s, and 299000 is subtracted from each value for ease of comparison. Based on this histogram, about how many observations were there between 750km/s and 800km/s? If this were a relative frequency histogram, what would be the height of that bar?

```
library(tidyverse)

lightspeed <- read_csv("lightspeed.csv")

ggplot(lightspeed,
       aes(x = Speed)) +
  geom_xxxxxxxxx(binwidth = XX,
                 boundary = 600,    ## this sets one boundary,
                                    ## and creates the rest of the bins
                                    ## using binwidth.
                 color = "black")
```

c. In R, construct side by side comparative boxplots of the measured speed of light for each of the five experiments. Compare the shape, center, and spread for each of the experiments.

```
ggplot(lightspeed,
       aes(x = XXXXX, y = XXXX, group = Expt)) +  ## fill in the blanks!!!
  geom_XXXXXXXXX() ## fill in the blanks
```

d. Now let’s drill down to experiment 3. Using the 1.5 IQR rule, would any of the values in this experiment be considered an outlier?

e. Next let’s drill down to experiment 5. Construct a histogram and boxplot for the experiment 5 data. Do the different graphical displays seem to say something different about the shape of the data? Explain the apparent contradiction and determine whether the experiment 5 data should be considered skewed or symmetrical.