Long Problem Set 2

1. In LPS01 you analyzed data on the concentration of NOX collected at a station along London's Marlybone Road. Reload the data and create four vectors: one for the spring months (March, April, and May), one for the summer months (June, July, and August), one for the fall months (September, October, and November), and one for the winter months (December, January, and February). For each discuss evidence for or against the claim that the data are normally distributed. See the schedule for August 29th if you need to download the data, which is in `LPS01.RData`.

First, we need to create our new set of vectors.

```
spring = c(mar.data$NOX, apr.data$NOX, may.data$NOX)
summer = c(jun.data$NOX, jul.data$NOX, aug.data$NOX)
fall = c(sep.data$NOX, oct.data$NOX, nov.data$NOX)
winter = c(dec.data$NOX, jan.data$NOX, feb.data$NOX)
```

Although it is tempting to begin by looking at a QQ plot (via qqnorm), a histogram, or a stripchart, we should consider first whether our data meet more general requirements for a normal distribution:

1. that results for NOX are continuous: sure, within the precision of our ability to measure NOX
2. that results for NOX are unbounded: well, no, at least at the lower end as the smallest possible concentration of NOX is 0 µg/m^3; however, this likely isn't a problem as the smallest recorded value of 8.6 µg/m^3 suggests that values approaching 0 µg/m^3 are rare
3. that the results for NOX are random and show no systematic pattern in, for this data, time: this is unclear without looking at the data, so let's plot the data for each season using a simple scatterplot of NOX levels as a function of each vector's index, as shown in Figure 1; note that the y-axis scales are kept constant to ensure that our interpretation of the plots is not affected by differences in scale.

```
old.par = par(mfrow = c(2, 2))
y.up = max(c(fall, spring, summer, winter), na.rm = TRUE)
plot(spring, type = "o", lwd = 2, cex = 0.5, col = "blue", main = "spring",
     ylim = c(0, y.up))
plot(summer, type = "o", lwd = 2, cex = 0.5,col = "blue", main = "summer",
     ylim = c(0, y.up))
plot(fall, type = "o", lwd = 2, cex = 0.5,col = "blue", main = "fall",
     ylim = c(0, y.up))
plot(winter, type = "o", lwd = 2, cex = 0.5,col = "blue", main = "winter",
     ylim = c(0, y.up))
```

```
par(old.par)
```

Our plots show evidence of what we might, for now, call a few "unusually" large NOX concentrations in winter, which we shouldn't take as implying that the underlying data is not normally distributed; all data is subject to such "unusual" values and there are ways to account for this. Setting aside these data points, we still see evidence of a pattern in the data for the winter of relatively low levels of NOX for the first half of the season and then an abrupt increase in values. The winter does not seem to pass our third criterion that the individual samples show no underlying systematic pattern.

The data for the remaining seasons seem to reflect random noise more than a distinct pattern (we might argue that the data for spring shows a downward tren in the concentration of NOX for the first half of the season, but the day-to-day variations are sufficiently large that this probably is not worth considering). For these, let's look at qqnorm plots, which are in Figure 2.

```
old.par = par(mfrow = c(1, 3))
qqnorm(summer, main = "summer", pch = 19, col = "blue", cex = 0.5)
qqline(summer, lwd = 2, col = "blue")
```

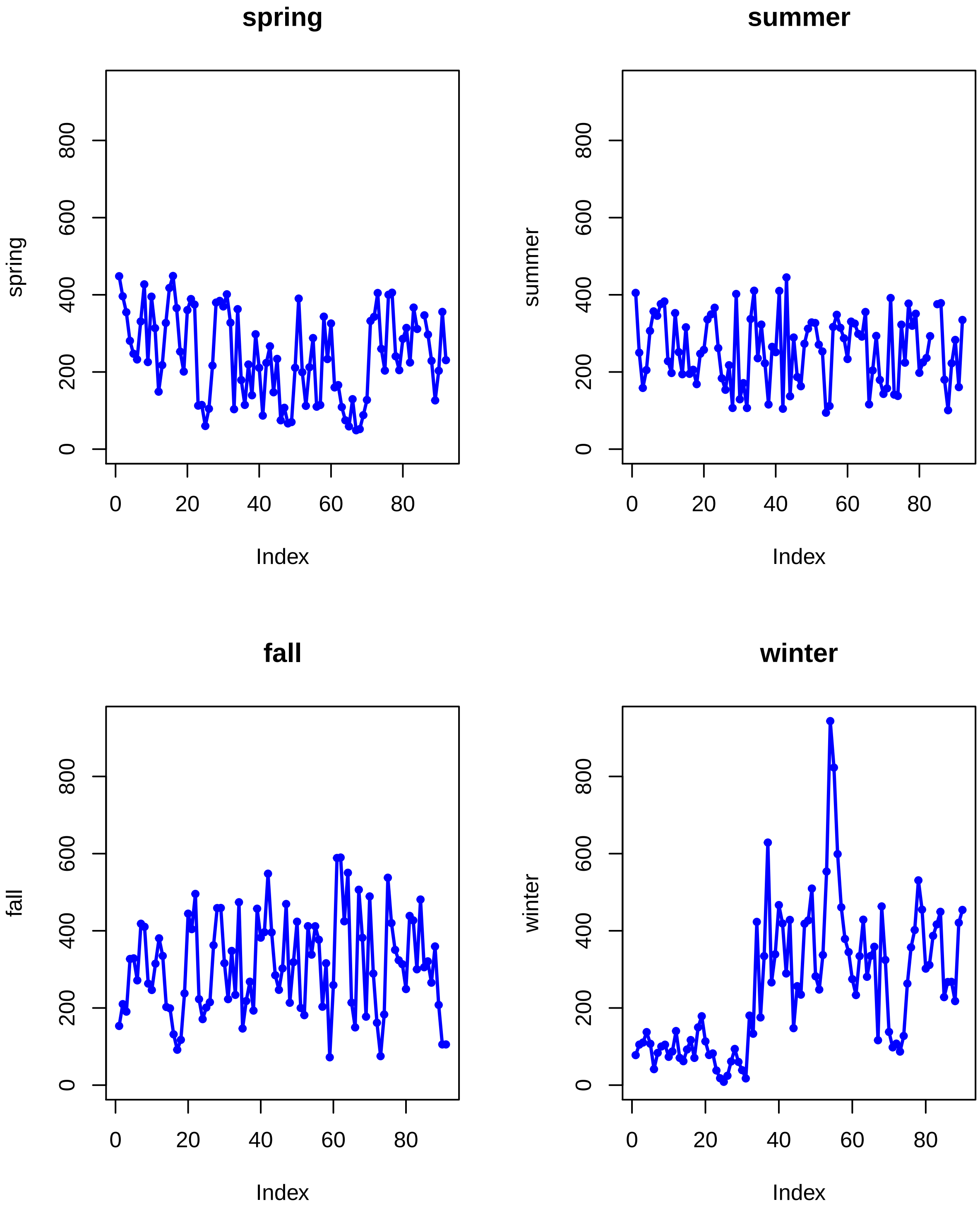

Figure 1: Scatterplots of Data by Season

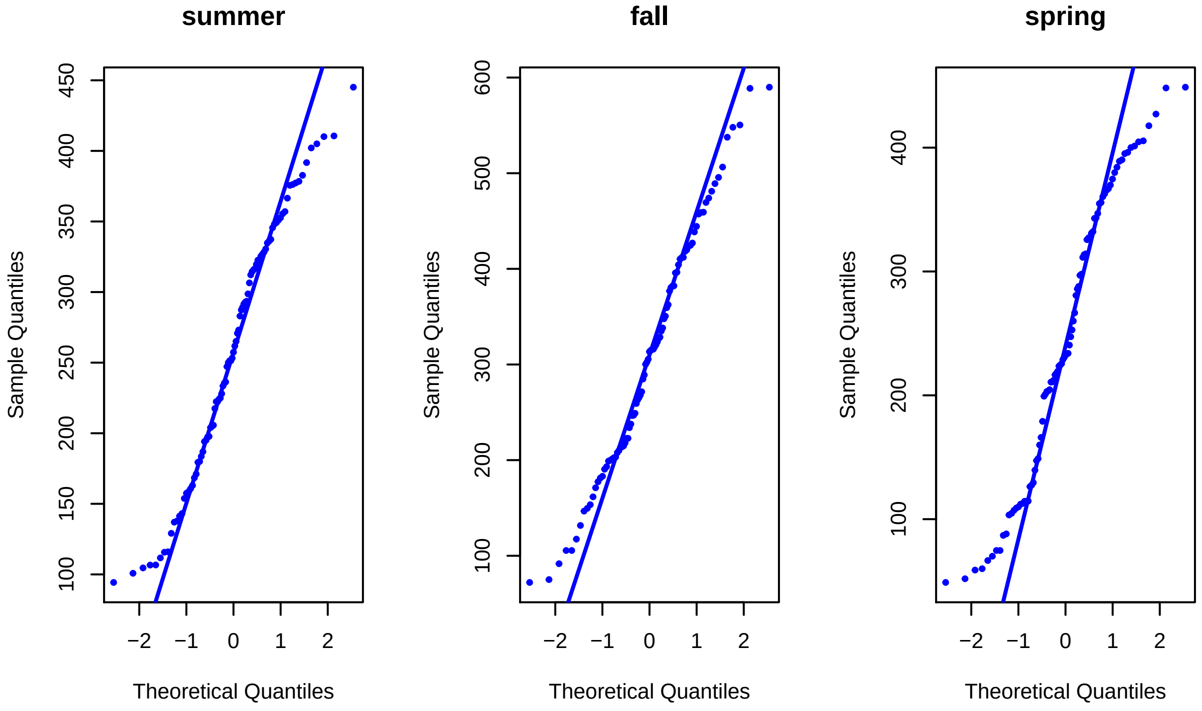

Figure 2: QQ Plots for Summer and Fall

```
qqnorm(fall, main = "fall", pch = 19, col = "blue", cex = 0.5)
qqline(fall, lwd = 2, col = "blue")
qqnorm(spring, main = "spring", pch = 19, col = "blue", cex = 0.5)
qqline(spring, lwd = 2, col = "blue")
```

```
par(old.par)
```

The qqnorm plot for all three seasons are reasonably consistent with a normal distribution, with the majority of points falling along the theoretical line for a normal distribution.

2. Using the data in the file `MM.RData` giving the colors of the candies in 30 1.69-oz bags of M&Ms, report the probability that a randomly selected bag will have (a) more than 18 brown M&Ms, (b) fewer than 12 red M&Ms, (c) between 10 and 15 blue M&Ms, and (d) either fewer than 5 or more than 10 orange M&Ms. You may assume that the experimental mean, $\bar{X}$, and experimental standard deviation, s, are appropriate estimates for the population's mean, μ, and standard deviation, σ. See the schedule for August 29th if you need to download the data, which is in `MM.RData`.

Figure 3 shows the areas under the normal distribution for all four scenarios.

```
old.par = par(mfrow = c(2, 2))
shadeArea(MMdata$brown, 18, 40, ylab = "frequency", xlab = "number of brown m&ms",
          col = "brown")
shadeArea(MMdata$red, 0, 12, ylab = "frequency", xlab = "number of red m&ms",
          col = "red")
shadeArea(MMdata$blue, 10, 15, ylab = "frequency", xlab = "number of blue m&ms",
          col = "blue")
shadeArea(MMdata$orange, 0, 5, ylab = "frequency", xlab = "number of orange m&ms",
          col = "orange")
shadeArea(MMdata$orange, 10, 40, overlay = TRUE, col = "orange")
```

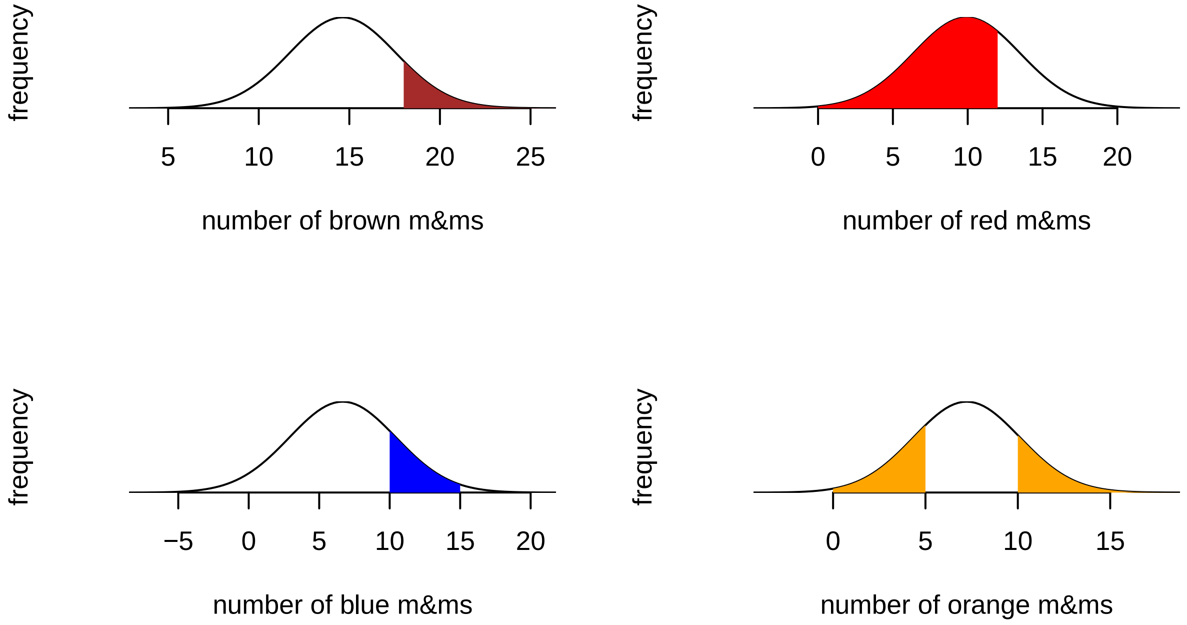

Figure 3: Scenarios For Problem 2

```
par(old.par)
```

The probabilities are found using the pnorm command; thus

```
prob.brown = pnorm(18, mean(MMdata$brown), sd(MMdata$brown), lower.tail = FALSE); prob.brown
```

```
## [1] 0.1262217
```

```
prob.red = pnorm(12, mean(MMdata$red), sd(MMdata$red), lower.tail = TRUE); prob.red
```

```
## [1] 0.7191311
```

```
prob.blue = pnorm(10, mean(MMdata$blue), sd(MMdata$blue), lower.tail = FALSE) -
  pnorm(15, mean(MMdata$blue), sd(MMdata$blue), lower.tail = FALSE); prob.blue
```

```
## [1] 0.1752505
```

```
prob.orange = pnorm(5, mean(MMdata$orange), sd(MMdata$orange), lower.tail = TRUE) +
  pnorm(10, mean(MMdata$orange), sd(MMdata$orange), lower.tail = FALSE); prob.orange
```

```
## [1] 0.3882071
```

3. The file `DistoNorm.RData` contains 26 vectors, each of which is a random sample of 100 values drawn from a uniform distribution with a minimum of 0 and a maximum of 1; these vectors are identified using the letters “a,”, “b,” … “z”. The file also contains a vector with the averages for each of the other 26 vectors. Partition the plot window into two rows of three columns each (see the document “Creating Plots Using R’s Base Graphics” for details on how to do this). Pick a four letter word that uses four different letters and plot histograms for the data in the vectors for each of your word’s letters. Next combine your four letters into a single vector—check your vector to ensure that it has 400 values—and plot its histogram. Finally, plot a histogram for the vector “avg”. Discuss how your results provide support for the central limit theorem.

Let’s use the word “chem” to explore the issues raised in this problem; the plots are shown in Figure 4; note that the histogram for avg includes a normal distribution curve based on its mean and standard deviation.

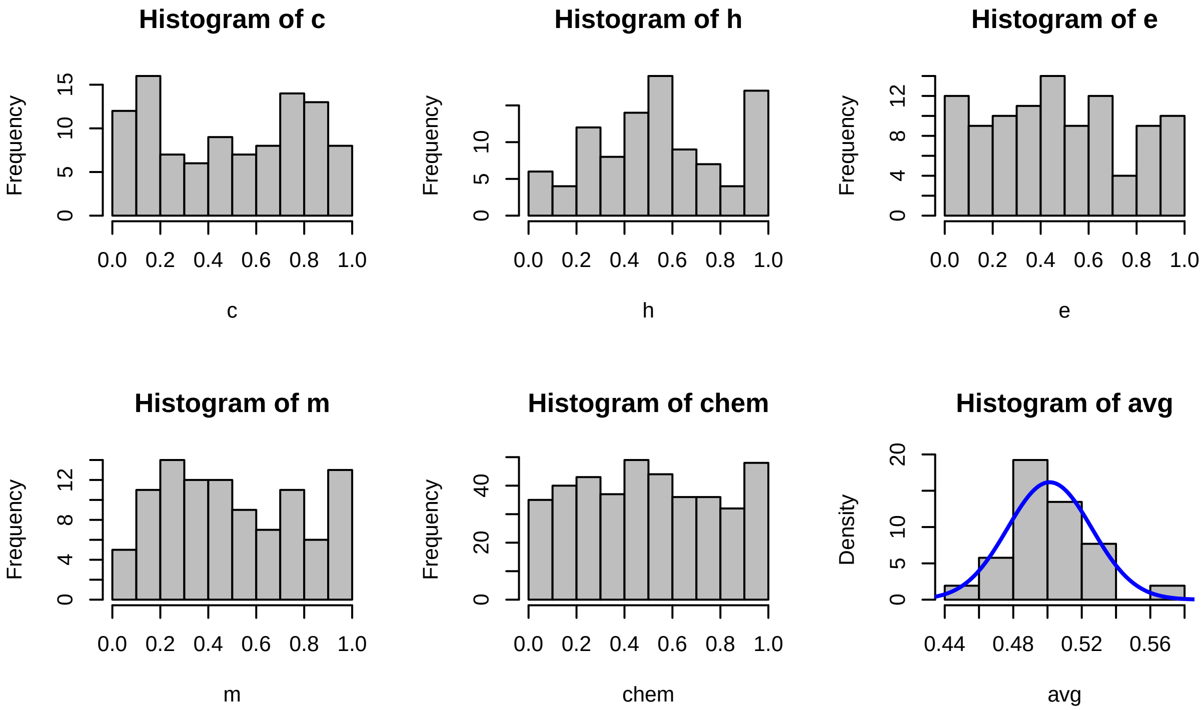

Figure 4: Results for Problem 3

```r
load("DisToNorm.RData")
old.par = par(mfrow = c(2,3))
chem = c(c, h, e, m)
hist(c, col = "grey", border = "black")
hist(h, col = "grey", border = "black")
hist(e, col = "grey", border = "black")
hist(m, col = "grey", border = "black")
hist(chem, col = "grey", border = "black")
hist(avg, col = "grey", border = "black", freq = FALSE)
x = seq(0.4, 0.6, 0.001)
lines(x, dnorm(x, mean(avg), sd(avg)), lwd = 2, col = "blue")
```

```r
par(old.par)
```

The histograms for objects c, h, e, and m are typical of a small sample drawn from a uniform distribution (and a sample of 100 is a small sample); although none of these histograms is particularly uniform in its distribution, each clearly is not representative of a normal distribution. The object chem, which has a sample of size 400, is sufficiently large that the underlying uniform distribution is more evident. The histogram for the object avg, which has just 26 elements, on the other hand, clearly shows the general shape of a normal distribution; this is consistent with the central limit theorem's claim that the distribution of means for samples drawn from any distribution will tend toward a normal distribution.

4. The script in the file `SimSample.R` defines the function `simsample`, which simulates the drawing of random samples from a parent population. The function takes four arguments:

   - `mean`: the true mean of the parent population
   - `stdev`: the true standard deviation of the parent population
   - `maxsize`: the largest sample to draw from the parent population; samples are drawn with all sizes from 1 to maxsize

- `reps`: the number of individual samples drawn for each possible sample size

  The function returns a plot that shows the mean for each sample drawn from the parent population—a total of maxsize × reps samples—a solid green line that marks the parent population's mean, and two dashed red lines that span the middle 50% of the parent population's values. An example of the code and the output is shown here

To discern how the maximum size and the number of replicates affects the function's output, we will explore two levels for each, running them in combinations of (low, low), (high, low), (low, high), and (high, high); Figure 5 shows typical results for the following conditions: (a) maxsize of 20 and reps of 20; (b) maxsize of 40 and reps of 20; (c) maxsize of 20 and reps of 40; and (d) maxsize of 40 and reps of 40.

```
source("SimSample.R")
old.par = par(mfrow = c(2, 2))
simsample(mean = 10, stdev = 1, maxsize = 20, reps = 20, main = "maxsize = 20, reps = 20")
simsample(mean = 10, stdev = 1, maxsize = 40, reps = 20, main = "maxsize = 40, reps = 20")
simsample(mean = 10, stdev = 1, maxsize = 20, reps = 40, main = "maxsize = 20, reps = 40")
simsample(mean = 10, stdev = 1, maxsize = 40, reps = 40, main = "maxsize = 40, reps = 40")
```

```
par(old.par)
```

Each point is the mean value for one trial of a given sample size. From these results we see that there is more variability in the reported mean values for smaller sample sizes than there are for larger sample sizes and that the number of replicates does not seem particularly important. We also see that this variability between individual mean values seems to "settle down" once the sample size is greater than around, say, 10; a somewhat larger sample size of, say, at least 15 gives some confidence that the mean of the samples will approximate closely the mean of the population from which the samples are drawn.

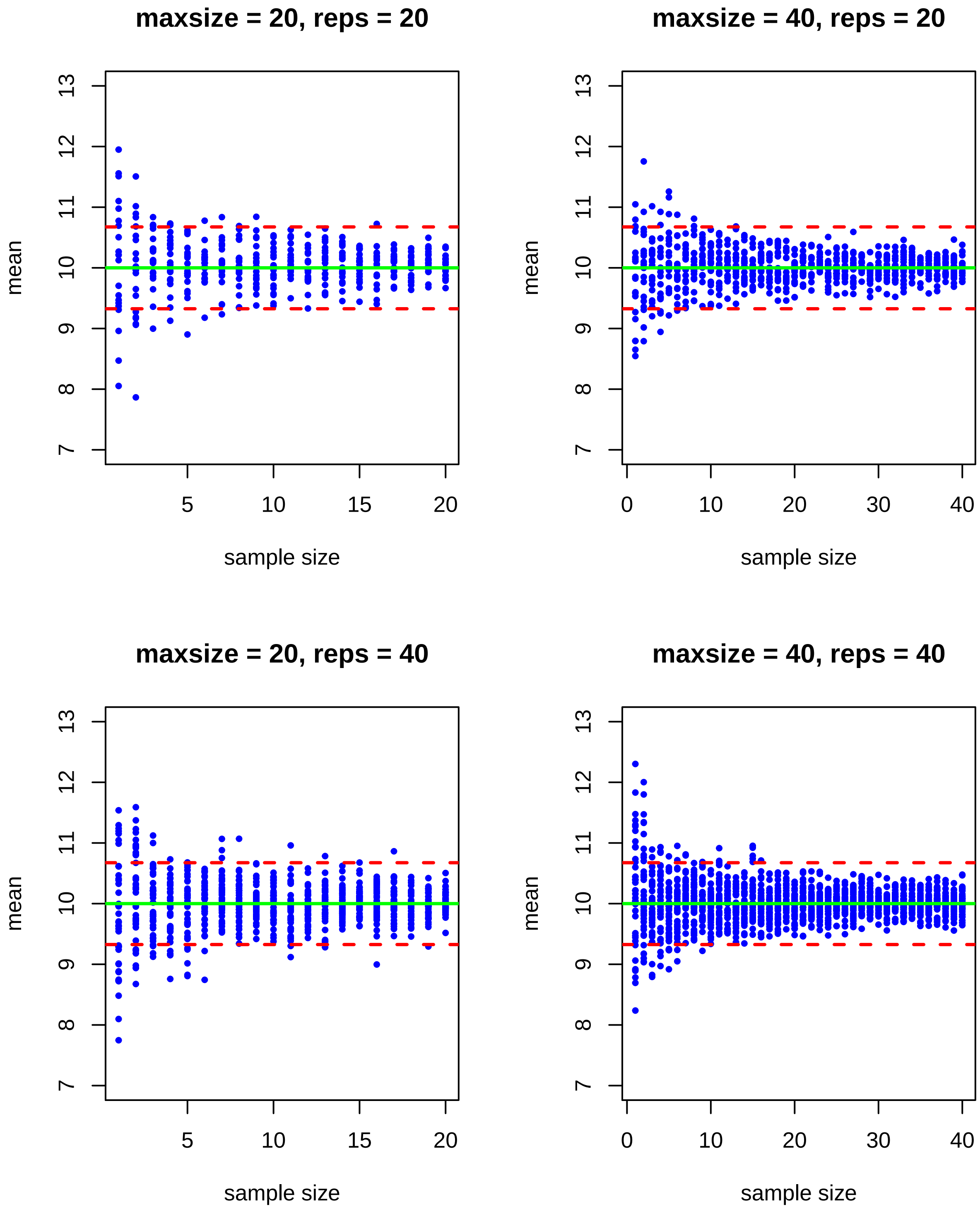

Figure 5: Results for Problem 4