Q4 Answers and Code

Yuxuan Zhang

```
nels<-dget("https://www2.stat.duke.edu/~pdh10/Teaching/610/Homework/nels_math_ses")
```

Part a

```
grand_mean <- mean(nels$mathscore, na.rm=TRUE)
grand_mean
```

```
## [1] 48.07446
```

```
group_mean <- aggregate(mathscore ~ school, data = nels, FUN = mean)
group_mean_var <- var(group_mean$mathscore)
group_mean_var
```

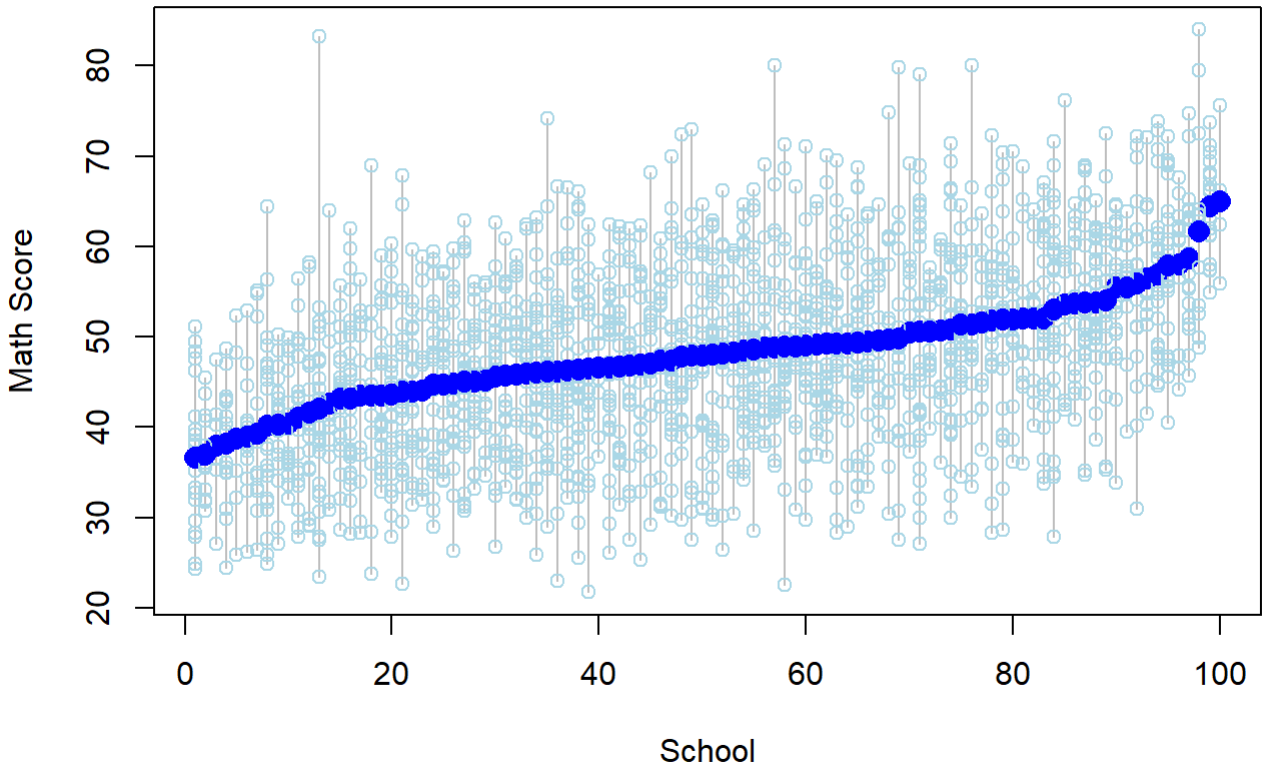
```
## [1] 30.99446
```

The grand mean is 48.074, the variance of the group means is 30.994.

Dotplot

```
gdotplot <- function(y, g, xlab="group", ylab="response", mcol="blue", ocol="lightblue", sortg
roups=TRUE, ...) {
  m <- length(unique(g))
  rg <- rank(tapply(y, g, mean), ties.method="first")
  if(sortgroups == FALSE) {
    rg <- 1:m
    names(rg) <- unique(g)
  }
  plot(c(1,m), range(y), type="n", xlab=xlab, ylab=ylab)

  for(j in unique(g)) {
    yj <- y[g == j]
    rj <- rg[match(as.character(j), names(rg))]
    nj <- length(yj)
    segments(rep(rj, nj), max(yj), rep(rj, nj), min(yj), col="gray")
    points(rep(rj, nj), yj, col=ocol, ...)
    points(rj, mean(yj), pch=16, cex=1.5, col=mcol)
  }
}
gdotplot(y = nels$mathscore, g = nels$school, xlab = "School", ylab = "Math Score", mcol = "blu
e", ocol = "lightblue")
```



Part b

```
result <- anova(lm(nels$mathscore ~ as.factor(nels$school)))
result
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
as.factor(nels$school)	99	48824.56	493.17737	5.834029	8.968176e-58
Residuals	1893	160024.02	84.53461	*NA*	*NA*

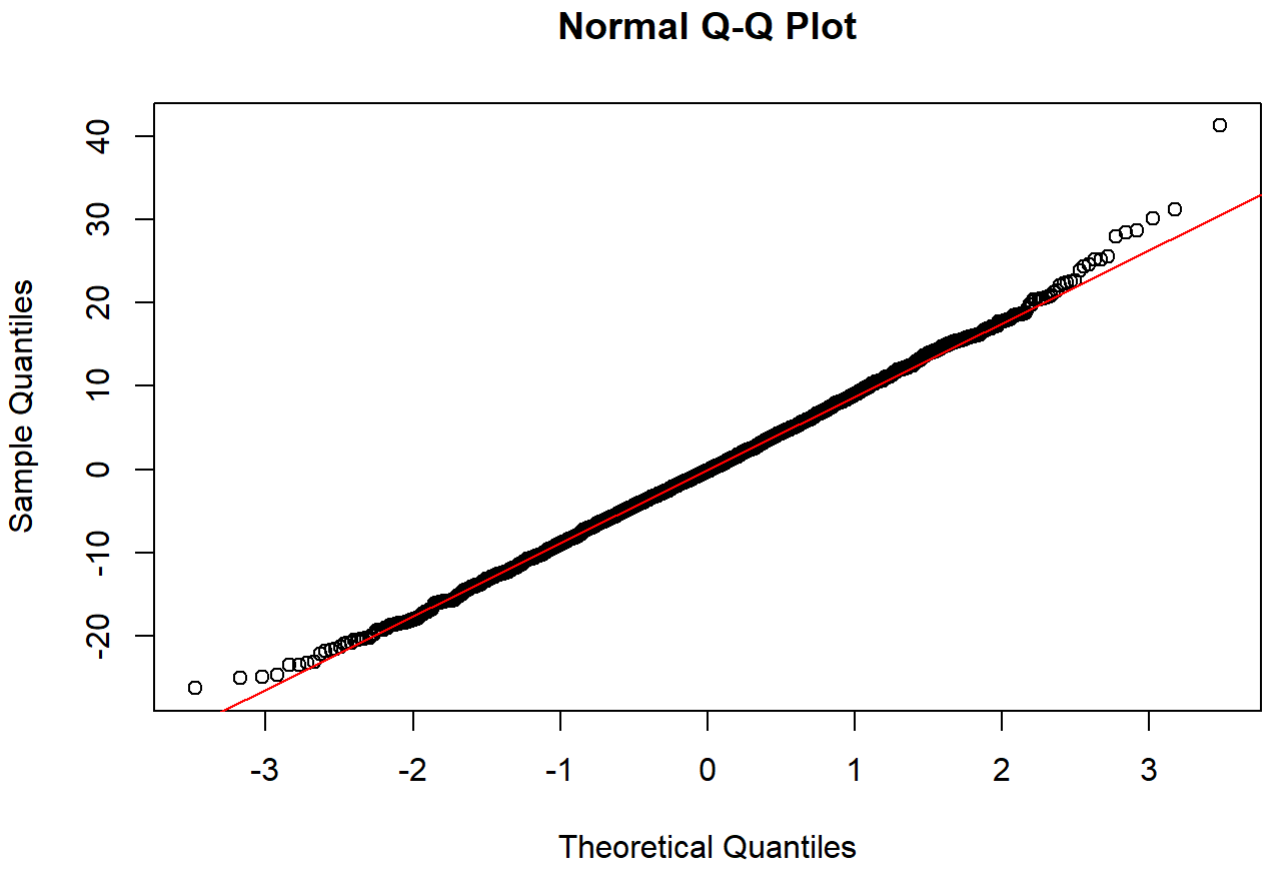
2 rows

The MSA is 493.18, much larger than MSW, which is 84.53. The F-score of 5.834 indicates that there is more variation between schools than within schools, suggesting that the schools are heterogeneous.

Our null hypothesis would be: All schools' mean math score are the same. In this case, out P-value is reported as less than 2.2e-16, which is extremely small. This suggest that there is strong evidence to reject the null hypothesis that all school means are equal. We can conclude that there are statistically significant differences in the mean math scores between schools.

Part c

```
# Compute residuals
model <- lm(nels$mathscore ~ as.factor(nels$school), data = nels)
residuals_model <- resid(model)
nels$residuals <- residuals_model
qqnorm(nels$residuals)
qqline(nels$residuals, col = "red")
```



The residuals follow a straight line on the Q-Q plot, which means that the normal model seem reasonable.

Part d

```
max_school <- group_mean[which.max(group_mean$mathscore), "school"]
min_school <- group_mean[which.min(group_mean$mathscore), "school"]
max_school
```

```
## [1] 3122
```

```
min_school
```

```
## [1] 1302
```

```
max_school_data <- subset(nels, school == max_school)$mathscore
min_school_data <- subset(nels, school == min_school)$mathscore
t.test(max_school_data)$conf.int
```

```
## [1] 51.88715 78.14785
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(min_school_data)$conf.int
```

```
## [1] 32.78774 40.37797
## attr(,"conf.level")
## [1] 0.95
```

```
length(max_school_data)
```

```
## [1] 4
```

```
length(min_school_data)
```

```
## [1] 21
```

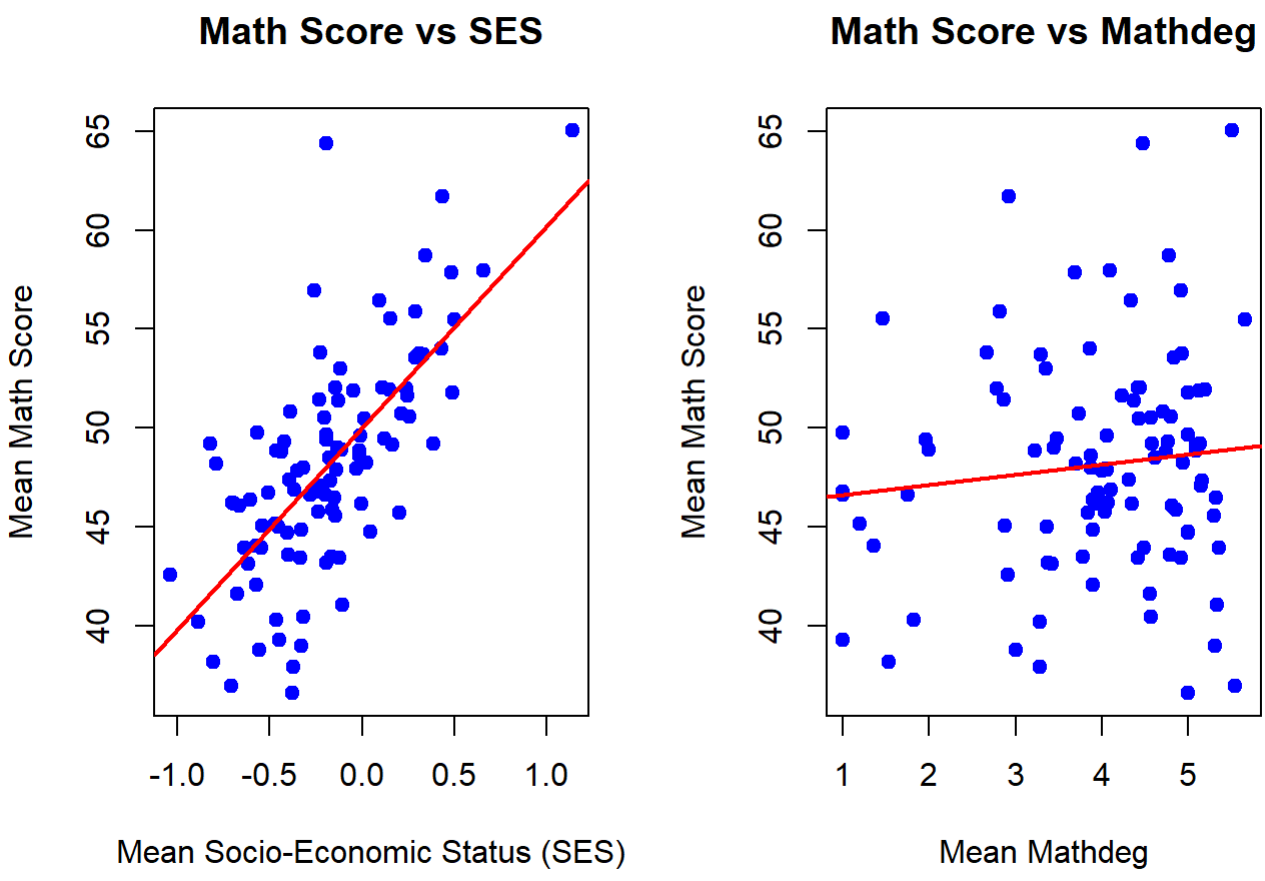
- The width of the highest mean school is $78.14785 - 51.88715 = 26.2607$.
- The width of the lowest mean school is $40.37797 - 32.78774 = 7.59023$.

We can see that the confidence interval width of the highest mean school is much larger than that of the lowest mean school. I believe the difference between the sample means of these two groups **does not** reflect the likely difference in subpopulation means. While the confidence intervals between the two schools does not overlap, showing a notable difference, this does not accurately reflect the true difference in the subpopulation means due to the sample sizes. We can see that the highest mean score school has only 4 samples, while the lowest mean score school has 21 samples. This causes the highest means school to have a large variance, and thus a large CI interval. Therefore, the difference may not accurately reflect the true difference in the subpopulation means.

Part e

```
group_means_mathdeg <- aggregate(mathdeg ~ school, data = nels, FUN = mean)
group_means_ses <- aggregate(ses ~ school, data = nels, FUN = mean)
group_means_mathscore <- aggregate(mathscore ~ school, data = nels, FUN = mean)
group_means_combined <- merge(group_means_mathscore, group_means_mathdeg, by = "school")
group_means_combined <- merge(group_means_combined, group_means_ses, by = "school")
par(mfrow = c(1, 2))
plot(group_means_combined$ses, group_means_combined$mathscore,
     xlab = "Mean Socio-Economic Status (SES)", ylab = "Mean Math Score",
     main = "Math Score vs SES", col = "blue", pch = 21, bg = "blue")
lm_ses <- lm(mathscore ~ ses, data = group_means_combined)
abline(lm_ses, col = "red", lwd = 2)

plot(group_means_combined$mathdeg, group_means_combined$mathscore,
     xlab = "Mean Mathdeg", ylab = "Mean Math Score",
     main = "Math Score vs Mathdeg", col = "blue", pch = 21, bg = "blue")
lm_mathdeg <- lm(mathscore ~ mathdeg, data = group_means_combined)
abline(lm_mathdeg, col = "red", lwd = 2)
```



In the “Math Score vs SES” plot, there seems to be a positive relationship between the mean socio-economic status and the mean math score. As SES increases, the average math score tend to increase as well. This suggests that students from schools with higher SES on average tend to perform better in math. This could indicate that socio-economic factors play a role in student performance. Higher SES could mean more resources, better access to learning materials, and perhaps higher quality teaching, all of which might contribute to higher average math scores.

In the “Math Score vs Mathdeg”, there doesn’t seem to be a clear linear trend between mean mathdeg. The points are more spread out, indicating a weaker or less clear relationship.

Based on our observations of the two plots, we conjecture that SES seems to be an important source of heterogeneity for the school-specific mean math score, as evidenced by the positive relationship in the first plot.