Math 255 - Homework 8

Colin Pi

Due in class, Friday May 17

Problem 1

Lohr textbook ch. 5 exercise 23

$$ICC = 1 - \frac{M}{M-1}\frac{SSW}{SSTot}$$
$$\frac{M}{M-1}\frac{SSW}{SSTot} = 1 - ICC$$
$$SSW = \frac{M-1}{M}SSTot(1-ICC)$$
$$MSW = \frac{SSW}{N(M-1)} = \frac{1}{N(M-1)}\frac{M-1}{M}SSTot(1-ICC) = \frac{SSTot}{NM}(1-ICC) = \frac{(NM-1)}{NM}S^2(1-ICC)$$

$$ICC = 1 - \frac{M}{M-1}\cdot\frac{SSW}{SSTot} = 1 - \frac{M}{M-1}\cdot\frac{SSTot-SSB}{SSTot} = 1 - \frac{M}{N-1}\cdot(1-\frac{SSB}{SSTot})$$
$$\frac{M}{M-1}\cdot\frac{SSB}{SSTot} = ICC - 1 + \frac{M}{M-1}$$
$$\frac{SSB}{SSTot} = \frac{M-1}{M}(ICC-1)+1$$
$$SSB = SSTot\cdot\frac{M-1}{M}(ICC-1)+SSTot$$
$$MSB = \frac{SSB}{N-1} = \frac{SSTot}{N-1}\cdot\frac{M-1}{M}(ICC-1)+\frac{SSTot}{N-1} = \frac{SSTot}{N-1}\frac{1}{M}((M-1)\cdot ICC - M + 1 + M) =$$
$$\frac{(NM-1)}{M(N-1)}S^2(1+(M-1)ICC)$$

Problem 2

Lohr textbook ch. 5 exercise 11. Use the following cluster-level data set:

```
> audit <- read.csv("http://math.carleton.edu/kstclair/data/audit.csv")
> audit$N <- 828
> audit$n <- 85
> audit$wts <- audit$N/audit$n
>
> audit.design <- svydesign(id = ~1, fpc = ~N, weights = ~wts,
+     data = audit)
```

(a)

$M = 215$. Therefore, we can use unbiased estimator to get the error rate because we know $M_0 = NM = 18275$ so that we can plug it into $\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{M_0}$ and $SE[\hat{\bar{y}}] = \frac{N}{M_0}SE[\bar{t}]$, where $\hat{t}_{unb} = \frac{N}{n}\sum_{i=1}^{n} t_i$, $SE[\bar{t}] = \sqrt{(1-\frac{n}{N})\frac{s_t^2}{n}}$
We can use ratio estimation to produce a same result as unbiased estimation because the cluster sizes are equal to each other.

```
> svyratio(~errors, ~fields, audit.design)
Ratio estimator: svyratio.survey.design2(~errors, ~fields, audit.design)
Ratios=
           fields
errors 0.002024624
SEs=
            fields
errors 0.0003570679
```

$\hat{p}_{unb} = \hat{p}_r = 0.002024624$, $SE[\hat{p}_{unb}] = SE[\hat{p}_r] = 0.0003570679$

(b)

```
> svytotal(~errors, audit.design)
        total     SE
errors 360.42 63.565
>
> 828 * mean(audit$errors)
[1] 360.4235
> 828 * sqrt((1 - 85/828) * var(audit$errors)/85)
[1] 63.56523
```

$$\hat{t}_{unb} = \frac{N}{n}\sum_{i=1}^{n} t_i = 360.42,\ SE[\hat{t}_{unb}] = N\sqrt{(1-\frac{n}{N})\frac{s_t^2}{n}} = 63.565$$

(c)

$\hat{p} = 0.002024624$

$$V[\hat{p}_{srs}] = (1-\frac{n}{N})\frac{\hat{p}(1-\hat{p})}{n-1} = (1-\frac{18,275}{178,020})\frac{0.002024624\cdot(1-0.002024624)}{18,275-1} = 9.921768e-08$$

$$V[\hat{p}_{clus}] = SE[\hat{p}_{clus}]^2 = 0.0003570679^2 = 1.274975e-07$$

$$DEff = \frac{V[\hat{p}_{clus}]}{V[\hat{p}_{clus}]} = \frac{1.274975e-07}{9.921768e-08} = 1.285028$$

```
> V_clus <- 0.0003570679^2
> V_srs <- (1 - 18275/178020) * 0.002024624 * (1 - 0.002024624)/18274
> Deff <- V_clus/V_srs
> Deff
[1] 1.285028
```

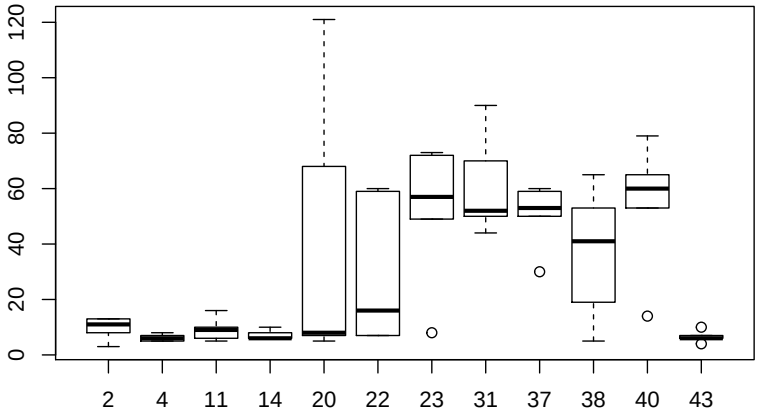
Problem 3

Lohr textbook ch. 5 exercise 10.

```
> books <- read.csv("http://math.carleton.edu/kstclair/data/books.csv")
```

(a)

```
> boxplot(replace ~ shelf, data = books)
```



The mean cost of the replacement varies significantly by the shelves. Also the variance differs considerably as well. The replacement cost of the books in shelf 2, 4, 11, 14, 37, 40, and 43 are way more homogenous than that of the books shelf 20, 22, and 38.

(b)

```
> books$elem.id <- 1:nrow(books)
> books <- books %>% group_by(shelf) %>% mutate(mi = n()) %>% ungroup()
> books$N <- 44
> books$wts <- (books$N * books$Mi)/(n_distinct(books$shelf) *
+     books$mi)
> books.design <- svydesign(id = ~shelf + elem.id, fpc = ~N + Mi,
+     weights = ~wts, data = books)
> svytotal(~replace, books.design)
        total     SE
replace 32638 5733.5
> cv <- 5733.5/32637.7
> cv
[1] 0.1756711
```

$$\hat{t}_{unbiased} = 32638,\ SE[\hat{t}_{unbiased}] = 5733.5,\ CV[\hat{t}_{unbiased}] = \frac{SE[\hat{t}_{unbiased}]}{\hat{t}_{unbiased}} = 0.1756711$$

(c)

Since we do not know M_0, we have to use ratio estimation.

$$\hat{\bar{y}}_r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{M_i}\frac{N}{n}\frac{M_i}{m_i}y_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{M_i}\frac{N}{n}\frac{M_i}{m_i}}$$

```
> svymean(~replace, books.design)
          mean     SE
replace 23.611 5.4759
> cv.2 <- 5.4759/23.611
> cv.2
[1] 0.2319216
```

$$\hat{\bar{y}}_r = 23.611,\ SE[\hat{\bar{y}}_r] = 5.4759,\ CV[\hat{\bar{y}}_r] = \frac{SE[\hat{\bar{y}}_r]}{\hat{\bar{y}}_r} = 0.2319216$$

Problem 4

Lohr textbook ch. 5 exercise 10.

```
> summary(aov(replace ~ as.factor(shelf), data = books))
                Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(shelf) 11  25571  2324.6   4.759 6.58e-05 ***
Residuals        48  23445   488.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> s2 <- (48 * 488.4 + 11 * 2324.6)/59
> r2a <- 1 - 488.4/s2
```

$$\hat{S}^2 = \frac{\hat{SSTO}}{dfTO} = \frac{\hat{SSW} + \hat{SSB}}{59} = \frac{(48)\hat{MSW} + (11)\hat{MSB}}{1319} + \frac{48 \cdot 488.4 + 11 \cdot 2324.6}{59} = 830.7423729$$

$$R_a^2 = 1 - \frac{\hat{MSW}}{\hat{S}^2} = 1 - \frac{488.4}{830.7424} = 0.4120921$$

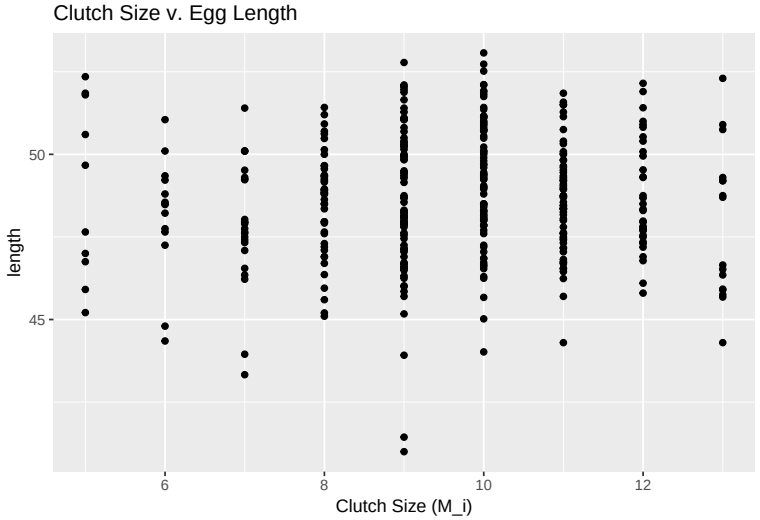
R_a^2 is less than 0.5. This suggest that the replacement cost of the books within the same shelf is not really homogeneous (maybe in a modest degree it is homogeneous).

$$m_{opt} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}} = \sqrt{\frac{10 \cdot 30(44-1)(1-0.1091834)}{4(1320-1)0.1091834}} = 6.3164535 \approx 7$$

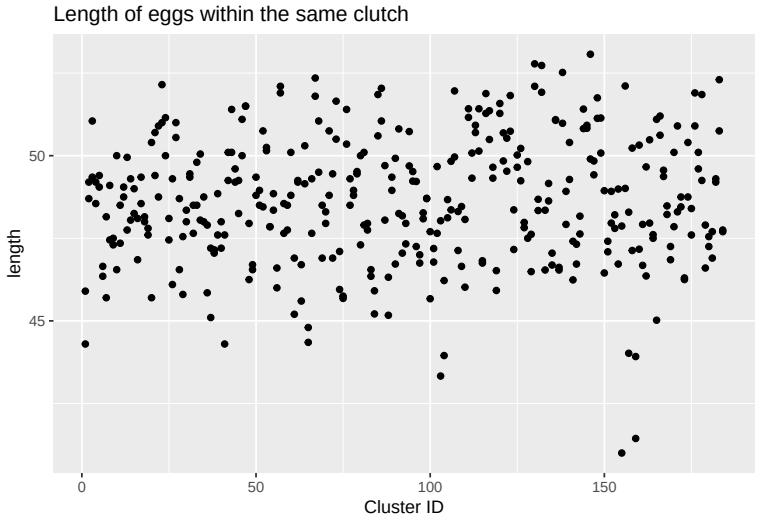
Problem 5

Lohr textbook ch. 5 exercise 12. Also use your "appropriate plot" to describe whether clusters look to be homogeneous with respect to length.

```
> coots <- read.csv("http://math.carleton.edu/kstclair/data/coots.csv")
>
> ggplot(coots, aes(x = csize, y = length)) + geom_point() + labs(x = "Clutch Size (M_i)",
+     title = "Clutch Size v. Egg Length")
```




```
> ggplot(coots, aes(x = clutch, y = length)) + geom_point() + labs(x = "Cluster ID",
+     title = "Length of eggs within the same clutch")
```



We cannot observe any strong homogeneity of the egg lengths within the clusters.

```
> coots$elem.id <- 1:nrow(coots)  ## unique id for each unique element (egg)
> coots <- coots %>% group_by(clutch) %>% mutate(mi = n()) %>%
+     ungroup()
> coots$wts <- coots$csize/coots$mi  ## since N is unknown, give relative weights Mi/mi
> coots.design <- svydesign(id = ~clutch + elem.id, weights = ~wts,
+     data = coots)
> mn.obj <- svymean(~length, coots.design)
> mn.obj
        mean     SE
length 48.65 0.1292
```

Since we do not have any information about M_0 (we cannot use $M_0 = NM$ because M is not equal)

$$\hat{\bar{y}}_r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{M_i}\frac{N}{n}\frac{M_i}{m_i}y_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{M_i}\frac{N}{n}\frac{M_i}{m_i}}$$

$\hat{\bar{y}}_r = 48.65$, $SE[\hat{\bar{y}}_r] = 0.1292$

Problem 6

Lohr textbook ch. 5 exercise 15.

```
> teachers.mi <- left_join(teachers, teachmi, by = "school")
> teachers.mi[, 3][teachers.mi[, 3] == -9] <- NA
> teachers.6 <- teachers.mi %>% filter(dist.x == "large")
```

(a)

If there is no full list of teachers in the study area, SRS might not be a viable option for collecting the data. And visiting the randomly selected clusters (schools) and conducting the survey to all the teachers in the school is much less costly and time-consuming that than visiting randomly selected teachers and conducting the survey. With an SRS of teachers, you might have to travel to a school just to conduct a survey to one teacher. Also, if the survey is collected in school level, confidentiality can be kept unless the respondent

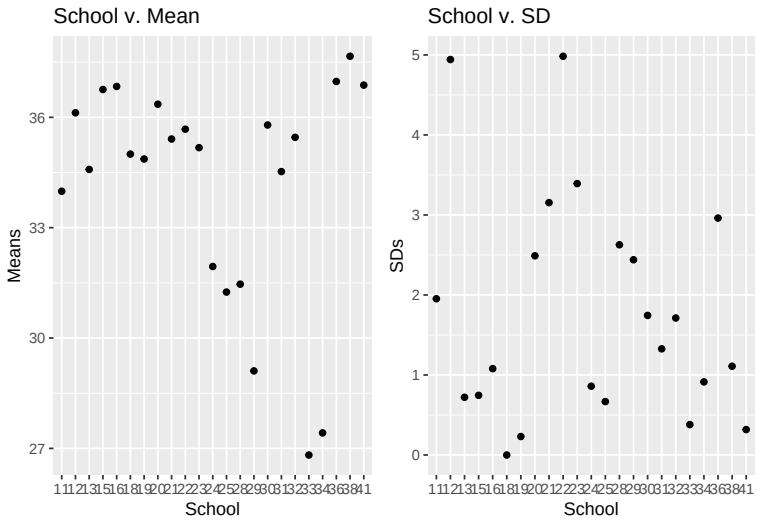
provides any personal information to the survey (researchers only know the school where the teacher is working at).

(b)

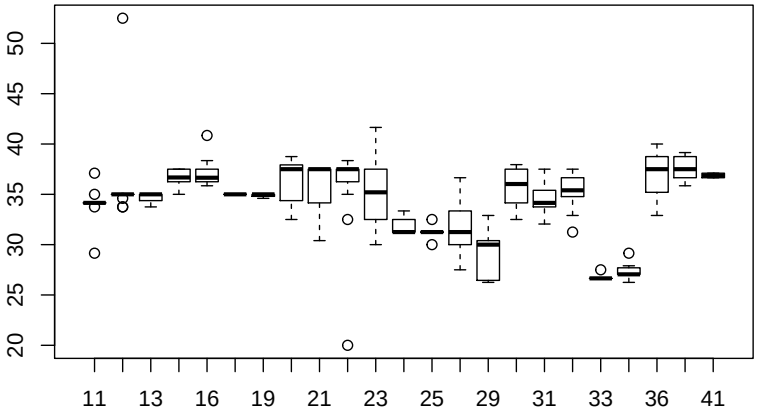
```
> hrwork <- teachers.6 %>% group_by(school) %>% summarize(means = mean(hrwork,
+     na.rm = TRUE), sds = sd(hrwork, na.rm = TRUE)) %>% ungroup()
>
> knitr::kable(hrwork)
```

school	means	sds
11	33.99000	1.9530318
12	36.12308	4.9424865
13	34.58333	0.7216878
15	36.75625	0.7460494
16	36.83958	1.0794503
18	35.00000	0.0000000
19	34.86667	0.2309401
20	36.35625	2.4894689
21	35.41000	3.1540450
22	35.67692	4.9834984
23	35.17500	3.3920495
24	31.94444	0.8600307
25	31.25000	0.6681531
28	31.46471	2.6273659
29	29.10625	2.4403509
30	35.79091	1.7448998
31	34.52500	1.3268993
32	35.45625	1.7122962
33	26.82000	0.3801316
34	27.42143	0.9137182
36	36.97500	2.9612779
38	37.66000	1.1095044
41	36.87500	0.3181981

```
>
> p1 <- ggplot(data = hrwork) + geom_point(aes(x = as.factor(school),
+     y = means)) + labs(x = "School", y = "Means", title = "School v. Mean")
>
> p2 <- ggplot(data = hrwork) + geom_point(aes(x = as.factor(school),
+     y = sds)) + labs(x = "School", y = "SDs", title = "School v. SD")
>
> grid.arrange(p1, p2, ncol = 2)
```



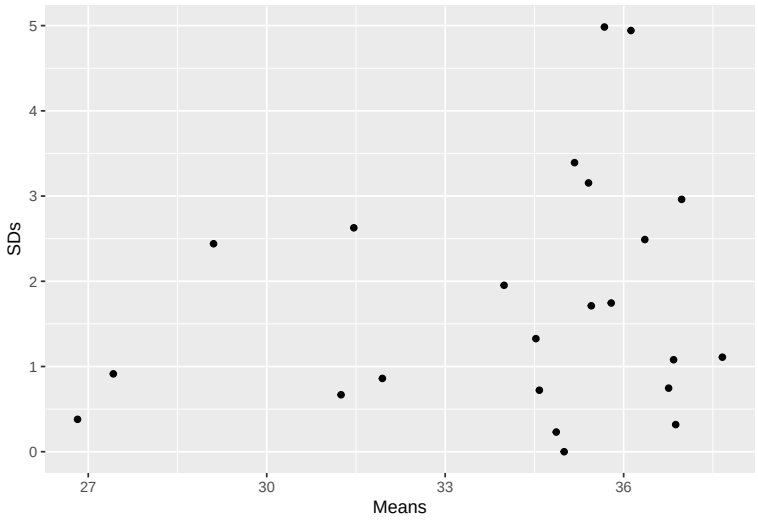

```
>
> boxplot(hrwork ~ as.factor(school), data = teachers.6)
```



The means of `hrwork` varies considerably by schools, whereas the standard deviations by school do not vary as much as the mean. This indicates that more variation is occurred from between the schools. I replaced -9 with `NA` and omitted them from the calculation.

(c)

```
> ggplot(data = hrwork) + geom_point(aes(x = means, y = sds)) +
+     labs(x = "Means", y = "SDs")
```



We do see there is a slight positive association between the means and the standard deviations. In other words, there is more variablity in schools with a higher workload.

(d)

```
> teachers.6$elem.id <- 1:nrow(teachers.6)
> teachers.6$wts <- teachers.6$popteach/teachers.6$ssteach
> teachers.clus <- svydesign(id = ~school + elem.id, weights = ~wts,
+     data = teachers.6)
> svymean(~hrwork, teachers.clus, na.rm = TRUE)
         mean     SE
hrwork 33.821 0.7444
```

$\hat{\bar{y}}_r = 33.821$, $SE[\hat{\bar{y}}_r] = 0.7444$

Problem 7

(a)

```
> response <- teachers.mi %>% group_by(school) %>% summarize(M = first(popteach),
+     m = first(ssteach))
> rate <- sum(response$m)/sum(response$M)
> rate
[1] 0.4111406
```

Response Rate $= \dfrac{\sum m_i}{\sum M_i}$ 0.4111406

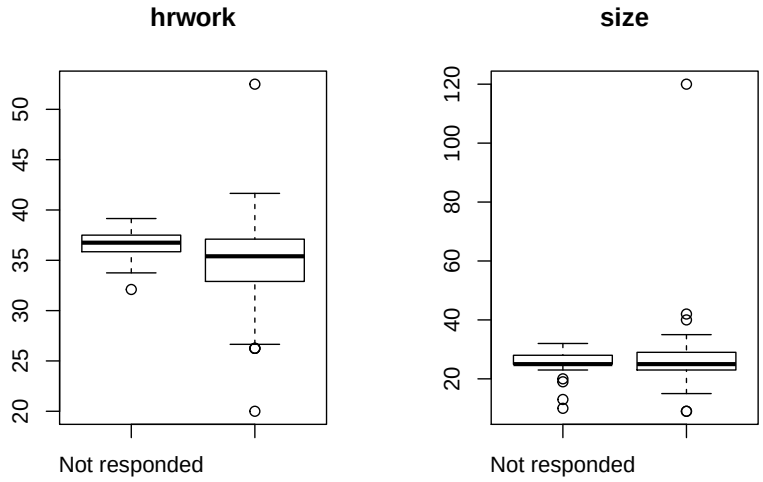
(b)

Teachers with heavier workload may not likely to respond to the survey because they may have no time to fill in the survey.

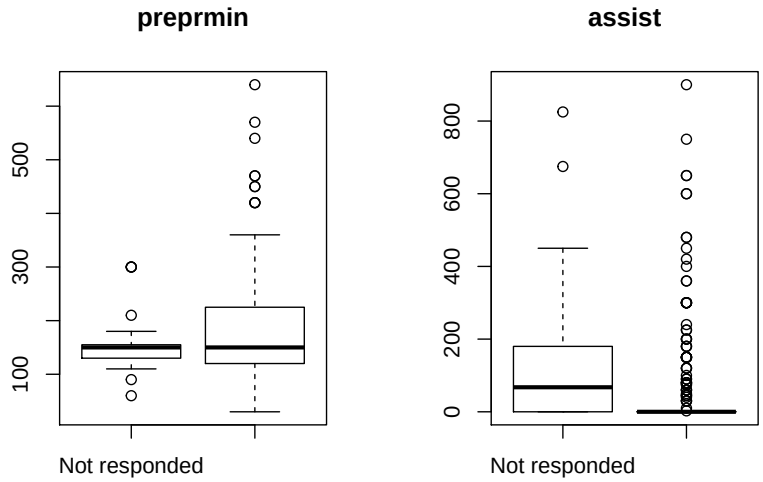
(c)

```
> summary(teachnr)
     hrwork           size          preprmin        assist
 Min.   :32.10   Min.   :10.00   Min.   : 60.0   Min.   :  0.0
 1st Qu.:35.95   1st Qu.:25.00   1st Qu.:130.0   1st Qu.:  0.0
 Median :36.75   Median :25.00   Median :150.0   Median : 67.5
 Mean   :36.46   Mean   :24.92   Mean   :160.2   Mean   :152.3
```

```
 3rd Qu.:37.50   3rd Qu.:28.00   3rd Qu.:153.8   3rd Qu.:172.5
 Max.   :39.15   Max.   :32.00   Max.   :300.0   Max.   :825.0
                 NA's   :1
> summary(teachers.6 %>% select(hrwork, size, preprmin, assist))
     hrwork           size           preprmin         assist
 Min.   :20.00   Min.   :  9.00   Min.   : 30.0   Min.   :  0.00
 1st Qu.:32.90   1st Qu.: 23.00   1st Qu.:120.0   1st Qu.:  0.00
 Median :35.40   Median : 25.00   Median :150.0   Median :  0.00
 Mean   :34.63   Mean   : 26.04   Mean   :170.8   Mean   : 55.13
 3rd Qu.:37.10   3rd Qu.: 29.00   3rd Qu.:225.0   3rd Qu.:  0.00
 Max.   :52.50   Max.   :120.00   Max.   :640.0   Max.   :900.00
 NA's   :3       NA's   :7        NA's   :14      NA's   :7
>
> teachnr$response = "Not responded"
> teachers.6$response = "Responded"
>
> comp <- rbind(teachers.6 %>% select(hrwork, size, preprmin, assist,
+     response), teachnr)
>
> par(mfrow = c(1, 2))
> boxplot(hrwork ~ response, data = comp, main = "hrwork")
> boxplot(size ~ response, data = comp, main = "size")
```




```
> par(mfrow = c(1, 1))
>
> par(mfrow = c(1, 2))
> boxplot(preprmin ~ response, data = comp, main = "preprmin")
> boxplot(assist ~ response, data = comp, main = "assist")
```




```
> par(mfrow = c(1, 1))
>
> svymean(~hrwork + size + preprmin + assist, teachers.clus, na.rm = TRUE)
            mean      SE
hrwork    33.864  0.7612
size      26.925  0.8152
preprmin 169.959  8.1437
assist    54.849 16.1937
> svymean(~hrwork, teachers.clus, na.rm = TRUE)
         mean     SE
hrwork 33.821 0.7444
```

In average, the minutes per week that a teacher's aide works with the teacher in the classroom is higher among the non-response group. Also the spread of this variable considerably bigger in nonresponse group than in original group.

(d)

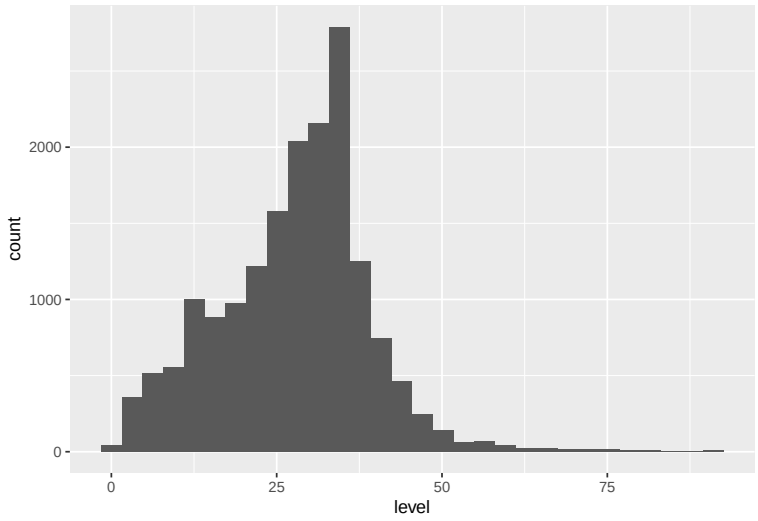
We do see there is a noticeable difference in the average minutes per week that a teacher's aide works with the teacher in the classroom between the original and the nonresponse group. Such difference indicates that there may be a nonresponse bias.

Problem 8

```
> ozone.long <- gather(ozone, key = hour, value = level, GMT1:GMT24)
```

(a)

```
> ggplot(ozone.long) + geom_histogram(aes(x = level), bins = 30)
Warning: Removed 265 rows containing non-finite values (stat_bin).
```

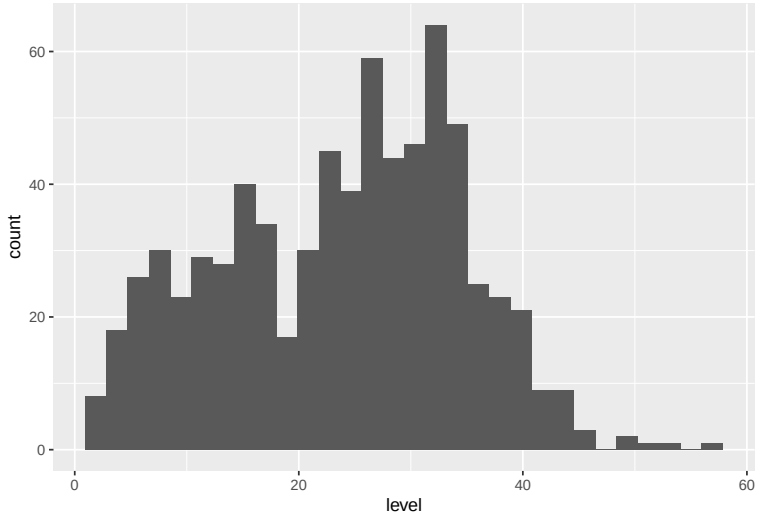



```
> mean(ozone.long$level, na.rm = T)
[1] 27.60979
> sd(ozone.long$level, na.rm = T)
[1] 11.42391
```

$\bar{y}_U = 27.60979$, $S_U = 11.42391$

(b)

```
> set.seed(70)
> k <- sample(1:24, 1)
> k
[1] 2
> gmt2 <- ozone.long %>% filter(hour == "GMT2")
>
> ggplot(gmt2) + geom_histogram(aes(x = level), bins = 30)
Warning: Removed 6 rows containing non-finite values (stat_bin).
```




```
> mean(gmt2$level, na.rm = T)
[1] 23.85497
```

Distribution is less skewed to the right and spread out than the population distribution. The mean is smaller than the population mean.

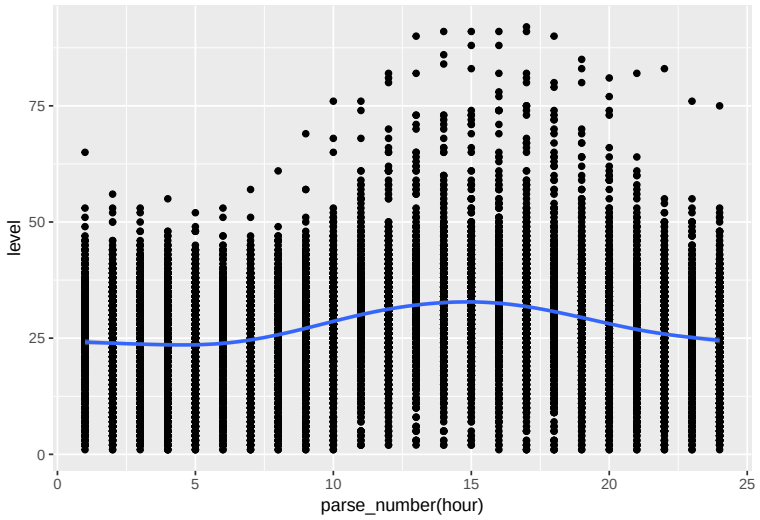
(c)

```
> gmt2$N <- nrow(ozone.long)
> gmt2$n <- nrow(gmt2)
> gmt2$wts <- gmt2$N/gmt2$n
> ozone.srs <- svydesign(ids = ~1, fpc = ~N, weights = ~wts, data = gmt2)
> svymean(~level, ozone.srs, na.rm = T)
        mean     SE
level 23.855 0.3917
> confint(svymean(~level, ozone.srs, na.rm = T))
         2.5 %   97.5 %
level 23.08723 24.62271
```

The CI does not include the population mean.

(d)

```
> ggplot(ozone.long, aes(x = parse_number(hour), y = level)) +
+     geom_point() + geom_smooth()
Warning: Removed 265 rows containing non-finite values (stat_smooth).
Warning: Removed 265 rows containing missing values (geom_point).
```



We can observe that both high and ozone levels were recorded in GMT2 period, so the ICC will be smaller than 0. Therefore, SRS SE may overestimate the true variability of the systematic sample.

(e)

```
> set.seed(70)
> start <- sample(1:96, size = 4, replace = FALSE)
> start
[1]  6 91 60 13
>
> rows.samp1 <- start[1] + 96 * (0:182)
> data.samp1 <- slice(ozone.long, rows.samp1)
> data.samp1$cluster <- "cluster 1"
> data.samp1$clustersize <- nrow(data.samp1)
>
> rows.samp2 <- start[2] + 96 * (0:182)
> data.samp2 <- slice(ozone.long, rows.samp2)
> data.samp2$cluster <- "cluster 2"
```

```
> data.samp2$clustersize <- nrow(data.samp2)
>
> rows.samp3 <- start[3] + 96 * (0:182)
> data.samp3 <- slice(ozone.long, rows.samp3)
> data.samp3$cluster <- "cluster 3"
> data.samp3$clustersize <- nrow(data.samp3)
>
> rows.samp4 <- start[4] + 96 * (0:182)
> data.samp4 <- slice(ozone.long, rows.samp4)
> data.samp4$cluster <- "cluster 4"
> data.samp4$clustersize <- nrow(data.samp4)
>
> data.sys <- bind_rows(data.samp1, data.samp2, data.samp3, data.samp4)
> data.sys$N <- 96
> data.sys$n <- n_distinct(data.sys$cluster)
> data.sys$wts <- data.sys$N/data.sys$n
>
> ozone.sys <- svydesign(id = ~cluster, fpc = ~N, weights = ~wts,
+     data = data.sys)
> confint(svymean(~level, ozone.sys, na.rm = T))
         2.5 %   97.5 %
level 27.72822 28.16315
```

The confidence interval still does not capture the true mean ozone level.