HUSTON Bootstrapping

Sara Huston

Contents

Assignment 1

Data Manipulation

Install and load all required packages (Don't install in rmd)

```
# install.packages('dplyr') install.packages('tidyverse')
# install.packages('gridExtra')
# install.packages('ggcorrplot')
# install.packages('formatR')

library("ggcorrplot")
library("gridExtra")
library("tidyverse")
library("dplyr")
library("formatR")
library(grid)
library(gridExtra)
```

Read in files based on type, headers, and with specific NA values accounted for

```
setwd("./Data")
w1_child <- read.csv("w1_child.csv", header = TRUE, na.strings = c("9",
    "8", "98", "99"))
w2_child <- read.table("w2_child.dat", header = TRUE, , na.strings = c("-999"))
w3_child <- read.csv("w3_child.csv", header = TRUE, na.strings = c("9",
    "8", "98", "99"))
w4_child <- read.csv("w4_child.csv", header = FALSE, na.strings = c("9",
    "8", "98", "99"))
names_w4_child <- read.table("names_w4_child.txt")
educinc <- read.csv("educinc.csv", header = TRUE, na.strings = c("."))
```

Add transposed variable names to w4_child based off of the provided text file

```
names(w4_child) <- t(names_w4_child)
```

Make all headers into lower case for easier merging

```
names(w1_child) <- tolower(names(w1_child))
names(w2_child) <- tolower(names(w2_child))
names(w3_child) <- tolower(names(w3_child))
names(w4_child) <- tolower(names(w4_child))
names(educinc) <- tolower(names(educinc))
```

Merge files by famid and select specified variables

```
w1234 <- (list(w1_child, w2_child, w3_child, w4_child, educinc) %>%
    reduce(full_join, by = "famid")) %>%
    dplyr::select(famid, c01cohort, c01gender, c01school, c01sibli,
        contains("atts"), contains("pcmp"), contains("attt"),
        contains("dscr"), contains("atod"), fameduc, income,
        c01sibli, contains("edex"))
```

Reverse code for pcmp 1 and 2

```
pcmp_01_02_cols <- c(grep("pcmp01", names(w1234)), grep("pcmp02",
    names(w1234)))
w1234[, pcmp_01_02_cols] <- 5 - w1234[, pcmp_01_02_cols]
```

Compute averages based on sets of columns for variable sets and place average into a new variable

```
w1234$c01attt <- rowMeans(w1234[c(grep("c01attt", names(w1234)))],
    na.rm = TRUE)
w1234$c04attt <- rowMeans(w1234[c(grep("c04attt", names(w1234)))],
    na.rm = TRUE)

w1234$c01pcmp <- rowMeans(w1234[c(grep("c01pcmp", names(w1234)))],
    na.rm = TRUE)
w1234$c04pcmp <- rowMeans(w1234[c(grep("c04pcmp", names(w1234)))],
    na.rm = TRUE)

w1234$c01dscr <- rowMeans(w1234[c("c01dscr07", "c01dscr08", "c01dscr09",
    "c01dscr10")], na.rm = TRUE)
w1234$c04dscr <- rowMeans(w1234[c("c04dscr07", "c04dscr08", "c04dscr09",
    "c04dscr10")], na.rm = TRUE)

w1234$c01atts <- rowMeans(w1234[c("c01atts03", "c01atts07", "c01atts08",
    "c01atts10")], na.rm = TRUE)
w1234$c02atts <- rowMeans(w1234[c("c02atts03", "c02atts07", "c02atts08",
    "c02atts10")], na.rm = TRUE)
w1234$c03atts <- rowMeans(w1234[c("c03atts03", "c03atts07", "c03atts08",
    "c03atts10")], na.rm = TRUE)
w1234$c04atts <- rowMeans(w1234[c("c01atts03", "c04atts07", "c04atts08",
    "c04atts10")], na.rm = TRUE)

w1234$c01atod <- rowSums(w1234[c("c01atod01", "c01atod02", "c01atod03",
    "c01atod04", "c01atod05", "c01atod06", "c01atod07", "c01atod08",
    "c01atod09")], na.rm = TRUE)
w1234$c04atod <- rowSums(w1234[c("c04atod01", "c04atod02", "c04atod03",
    "c04atod04", "c04atod05", "c04atod06", "c04atod07", "c04atod08",
    "c04atod09")], na.rm = TRUE)
```

Quiz Questions

1) Dimension of your final data frame (row x column)

```
dim(w1234)
```

```
## [1] 674 339
```

2) What is the max average education level obtained by parents ("fameduc")?

```
max_educ <- max(w1234$fameduc, na.rm = TRUE)
max_educ
```

```
## [1] 19
```

3) How many children have parents with this level of education?

```
sum(w1234$fameduc >= max_educ, na.rm = TRUE)
```

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

4) What is the mean for variable "c01attt" for those who have 4 siblings?

```
four_sibs <- w1234[w1234$c01sibli == 4 & !is.na(w1234$c01sibli),
    ]
mean(four_sibs$c01attt, na.rm = T)
```

```
## [1] 2.9073
```

5) Create difference scores between Waves 1 and 4 for all scales. Which scale has the greatest average absolute change (i.e., difference)?

```
score_variables <- c("atts", "pcmp", "attt", "dscr", "atod")
w1234[paste("difference_", score_variables, sep = "")] <- w1234[paste("c04",
    score_variables, sep = "")] - w1234[paste("c01", score_variables,
    sep = "")]
ave_abs_change <- abs(colMeans(w1234[, c("difference_atts", "difference_pcmp",
    "difference_attt", "difference_dscr", "difference_atod")],
    na.rm = TRUE))
ave_abs_change
```

```
## difference_atts difference_pcmp difference_attt difference_dscr difference_atod
##         0.14359         0.08697         0.48535         0.16869         9.62760
```

```
ave_abs_change[which.max(ave_abs_change)]
```

```
## difference_atod
##          9.6276
```

```
abs(mean(w1234$c01atts - w1234$c04atts, na.rm = TRUE))
```

```
## [1] 0.14359
```

```
abs(mean(w1234$c01pcmp - w1234$c04pcmp, na.rm = TRUE))
```

```
## [1] 0.08697
```

```
abs(mean(w1234$c01attt - w1234$c04attt, na.rm = TRUE))
```

```
## [1] 0.48535
```

```r
abs(mean(w1234$c01dscr - w1234$c04dscr, na.rm = TRUE))
```

```
## [1] 0.16869
```

```r
abs(mean(w1234$c01atod - w1234$c04atod, na.rm = TRUE))
```

```
## [1] 9.6276
```

Assignment 2

Plots

Plot 1

```r
w1234$gender_r[w1234$c01gender == 1] <- "female"
w1234$gender_r[w1234$c01gender == 2] <- "male"
```

Recode gender variables

```r
w1234$new_edu[w1234$fameduc < 7 & !is.na(w1234$fameduc)] <- "Elementary"
w1234$new_edu[w1234$fameduc >= 7 & w1234$fameduc < 13 & !is.na(w1234$fameduc)] <- "High School"
w1234$new_edu[w1234$fameduc >= 13 & w1234$fameduc < 17 & !is.na(w1234$fameduc)] <- "College"
w1234$new_edu[w1234$fameduc >= 17 & !is.na(w1234$fameduc)] <- "More than College"
```

Create a new variable called "newedu" w/ various levels of education

```r
w1234$new_edu <- factor(c(w1234$new_edu), levels = c("Elementary",
    "High School", "College", "More than College"), exclude = NA)
```

Order the education levels and factor

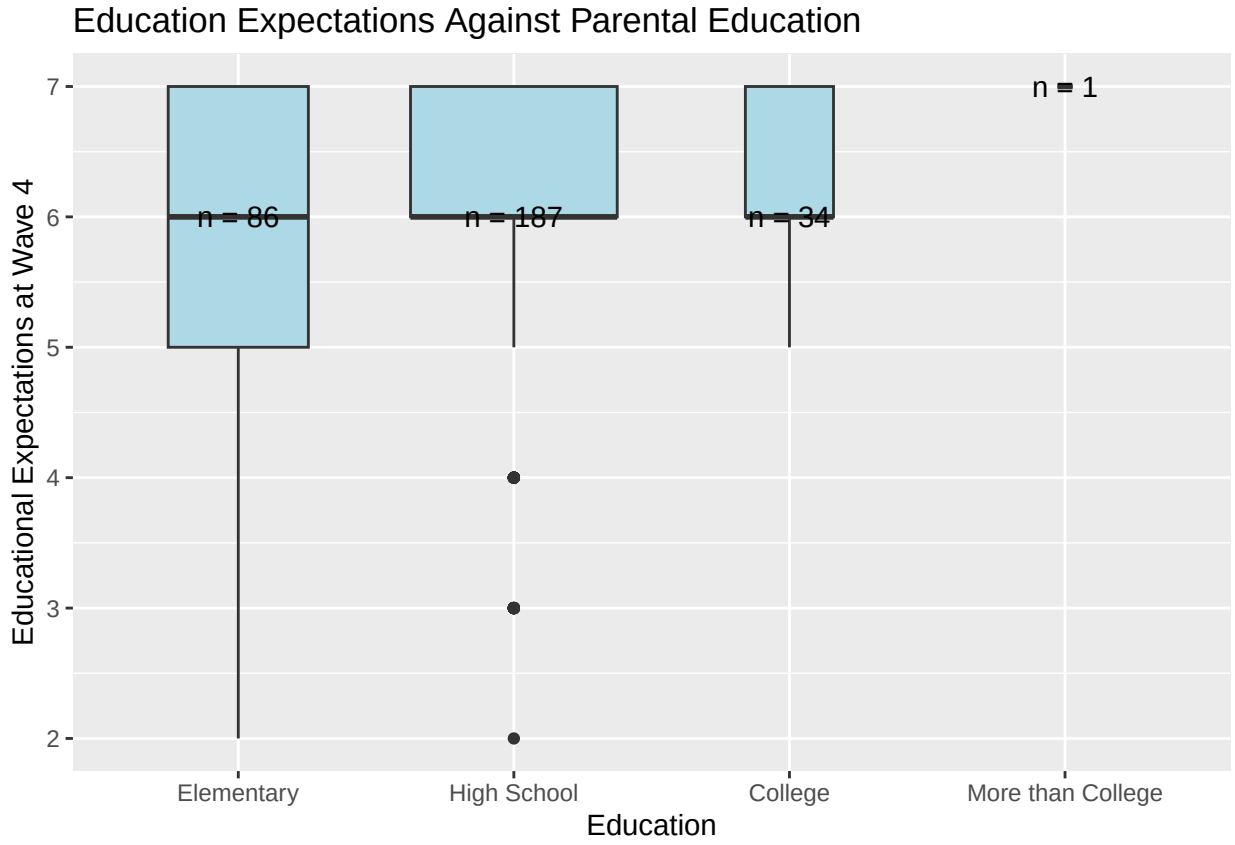
```r
w1234_clean <- subset(w1234, !is.na(new_edu) & !is.na(c04edex01) &
    !is.na(fameduc))
```

Remove na from each plot

```r
# function for number of observations
n_fun <- function(x) {
    return(data.frame(y = median(x), label = paste0("n = ", sum(!is.na(x)))))
}

g <- ggplot(data = w1234_clean, mapping = aes(new_edu, c04edex01))
g + geom_boxplot(data = w1234_clean, mapping = aes(new_edu, c04edex01),
    varwidth = T, fill = "lightblue") + stat_summary(fun.data = n_fun,
    geom = "text") + labs(title = "Education Expectations Against Parental Education",
    x = "Education", y = "Educational Expectations at Wave 4")
```

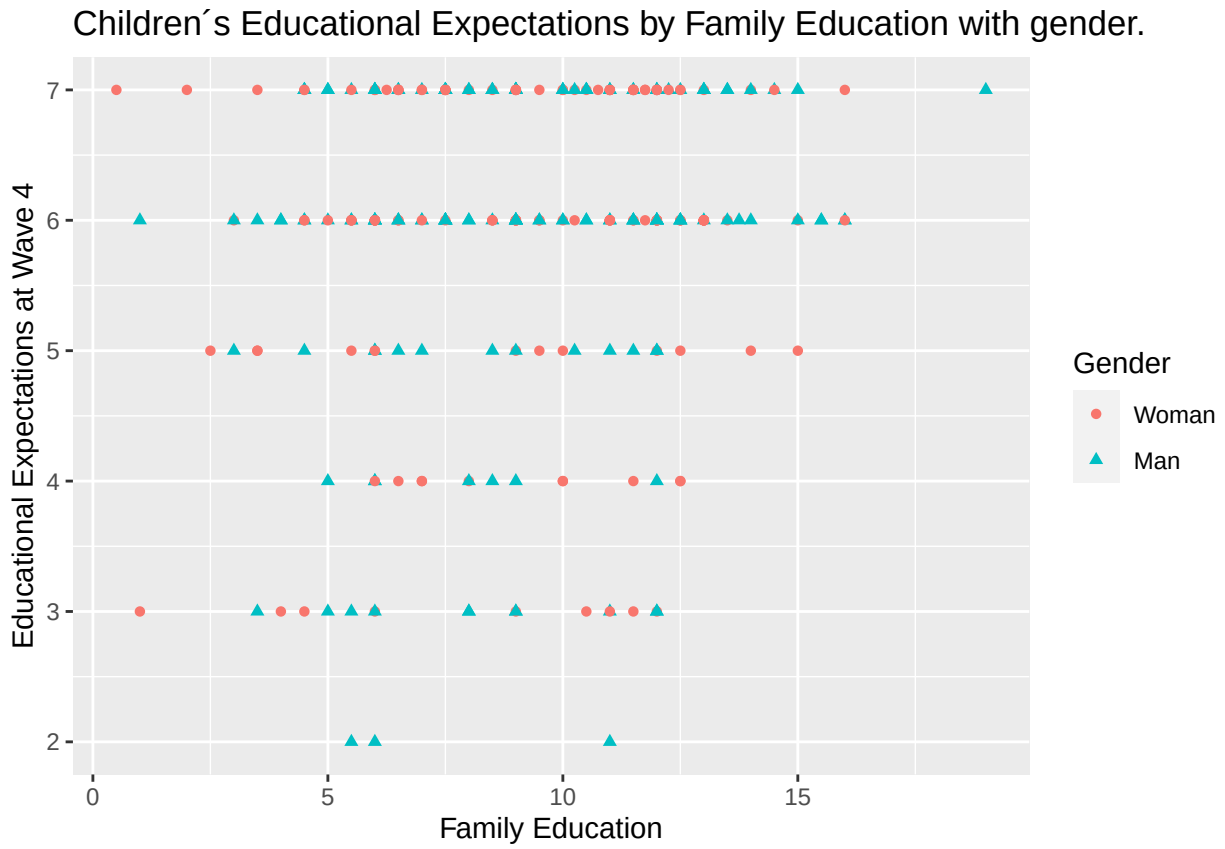
Create a box plot with education levels and educational experiences at wave 4



Plot 2

```r
ggplot(data = w1234_clean, mapping = aes(x = fameduc, y = c04edex01,
    color = gender_r, shape = gender_r)) + geom_point() + labs(title = "Children's Educational Expectati
    x = "Family Education", y = "Educational Expectations at Wave 4") +
    scale_shape_discrete(name = "Gender", breaks = c("female",
        "male"), labels = c("Woman", "Man")) + scale_colour_discrete(name = "Gender",
    breaks = c("female", "male"), labels = c("Woman", "Man"))
```

Create a scatter plot with educational expectations at wave 4 (c04edex01) on the Y-axis with "fameduc" on the X axis



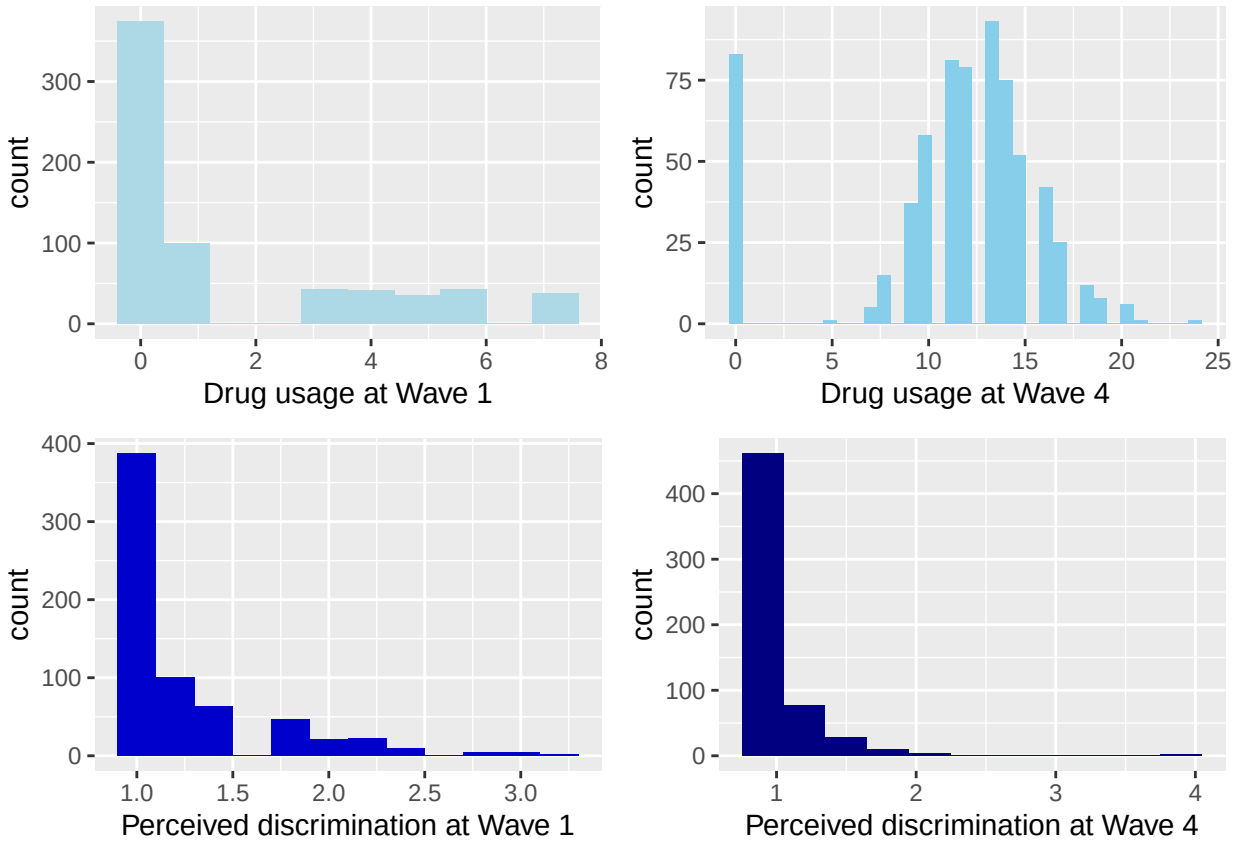
Plot 3

```
t1 <- ggplot(data = subset(w1234, !is.na(c01atod)), mapping = aes(c01atod))
t2 <- ggplot(data = subset(w1234, !is.na(c04atod)), mapping = aes(c04atod))
t3 <- ggplot(data = subset(w1234, !is.na(c01dscr)), mapping = aes(c01dscr))
t4 <- ggplot(data = subset(w1234, !is.na(c04dscr)), mapping = aes(c04dscr))

g1 <- t1 + geom_histogram(fill = "lightblue", binwidth = 0.8) +
    labs(x = "Drug usage at Wave 1")
g2 <- t2 + geom_histogram(fill = "skyblue", binwidth = 0.7) +
    labs(x = "Drug usage at Wave 4")
g3 <- t3 + geom_histogram(fill = "mediumblue", binwidth = 0.2) +
    labs(x = "Perceived discrimination at Wave 1")
g4 <- t4 + geom_histogram(fill = "navyblue", binwidth = 0.3) +
    labs(x = "Perceived discrimination at Wave 4")

grid.arrange(g1, g2, g3, g4, ncol = 2, nrow = 2)
```

Plot histograms of perceived discrimination and use of alcohol, tobacco, or other drugs at waves 1 and 4



Describe the distributions in two sentences:

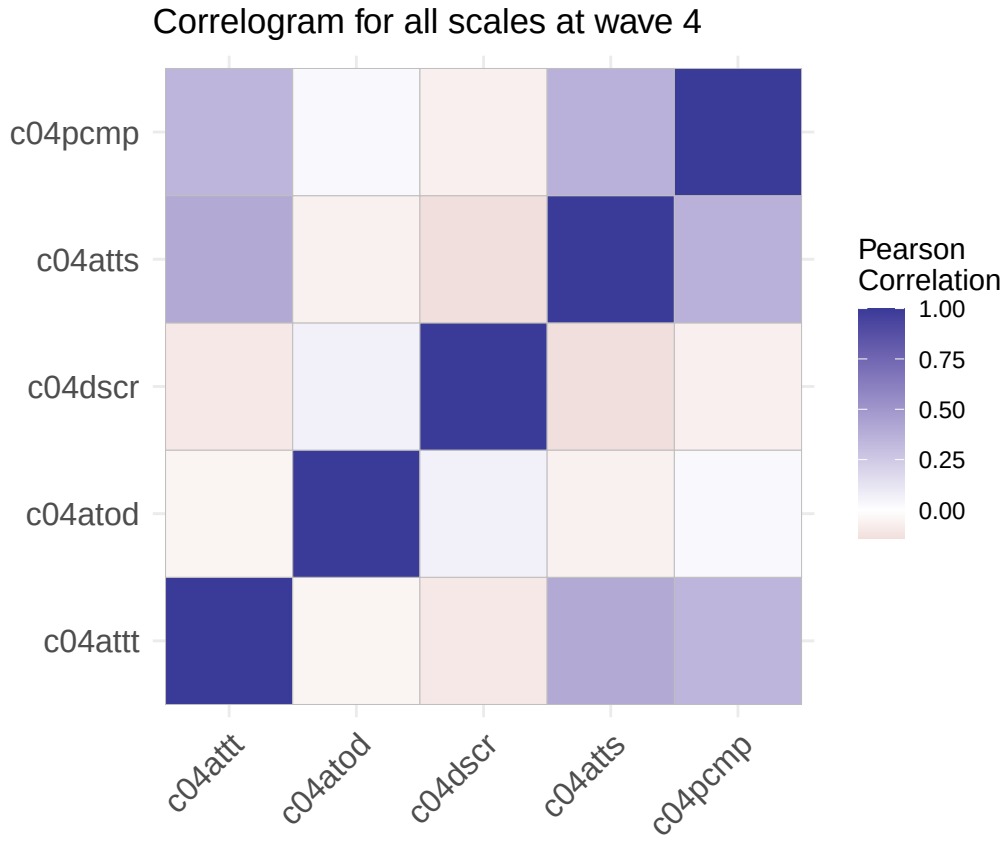
- Perceived discrimination decreases from wave 1 to wave 4, with an increase of ~100 in 1 scores.
- Alcohol, tobacco, or other drug use increased heavily from wave 1 to wave 4.

Plot 4

```
cor_scales <- cor(w1234[, c("c04attt", "c04atod", "c04dscr",
    "c04atts", "c04pcmp")], use = "complete.obs")
cor_scales <- round(cor_scales, 2)
ggcorrplot(cor_scales) + labs(title = "Correlogram for all scales at wave 4") +
    scale_fill_gradient2(name = "Pearson\nCorrelation")
```

Create a correlogram for all 5 scales at Wave 4

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```



Report in words which variables seem to be correlated.

- Child's peer competence (pcmp) seems to be correlated with child's attachment to teachers (attt) and child's attachment to school (atts). Child's attachment to teachers (attt) and child's attachement to school (atts) appear also correlated.

Analysis Questions

1) Conduct 2 tests of mean differences

```
hs_group <- subset(w1234, new_edu == "High School" & !is.na(c04edex01))
college_group <- subset(w1234, new_edu == "College" & !is.na(c04edex01))
t.test(hs_group$c04edex01, college_group$c04edex01, parid = TRUE)
```

```
##
##  Welch Two Sample t-test
##
## data:  hs_group$c04edex01 and college_group$c04edex01
## t = -2.47, df = 90.7, p-value = 0.015
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.593165 -0.064589
## sample estimates:
```

```
## mean of x mean of y
##    5.9947    6.3235
```

```
t.test(w1234$c01atod, w1234$c04atod, paird = TRUE)
```

```
##
##  Welch Two Sample t-test
##
## data:  w1234$c01atod and w1234$c04atod
## t = -46, df = 962, p-value <2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.0384  -9.2168
## sample estimates:
## mean of x mean of y
##    1.6202   11.2478
```

There was a significant difference in educational expectations at wave 4 between those whose parents have a high school education and those whos parents have a college education. There was a significant difference on alcohol, tobacco, and other drug use between wave 1 and wave 4.

2) Two tests for significant correlations

```
cor.test(w1234$c04attt, w1234$c04pcmp, na.action = na.omit)
```

```
##
##  Pearson's product-moment correlation
##
## data:  w1234$c04attt and w1234$c04pcmp
## t = 8.86, df = 588, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.27006 0.41258
## sample estimates:
##     cor
## 0.34329
```

```
cor.test(w1234$c04dscr, w1234$c04atod, na.action = na.omit)
```

```
##
##  Pearson's product-moment correlation
##
## data:  w1234$c04dscr and w1234$c04atod
## t = 1.62, df = 583, p-value = 0.11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.014216  0.147193
## sample estimates:
##      cor
## 0.066926
```

There was a significant difference at wave 4 between child's peer competence and child's attachment to teachers. There was not a significant difference at wave 4 between child's alcohol, tobacco, and other drug use and child's perceived discrimination.

Assignment 4

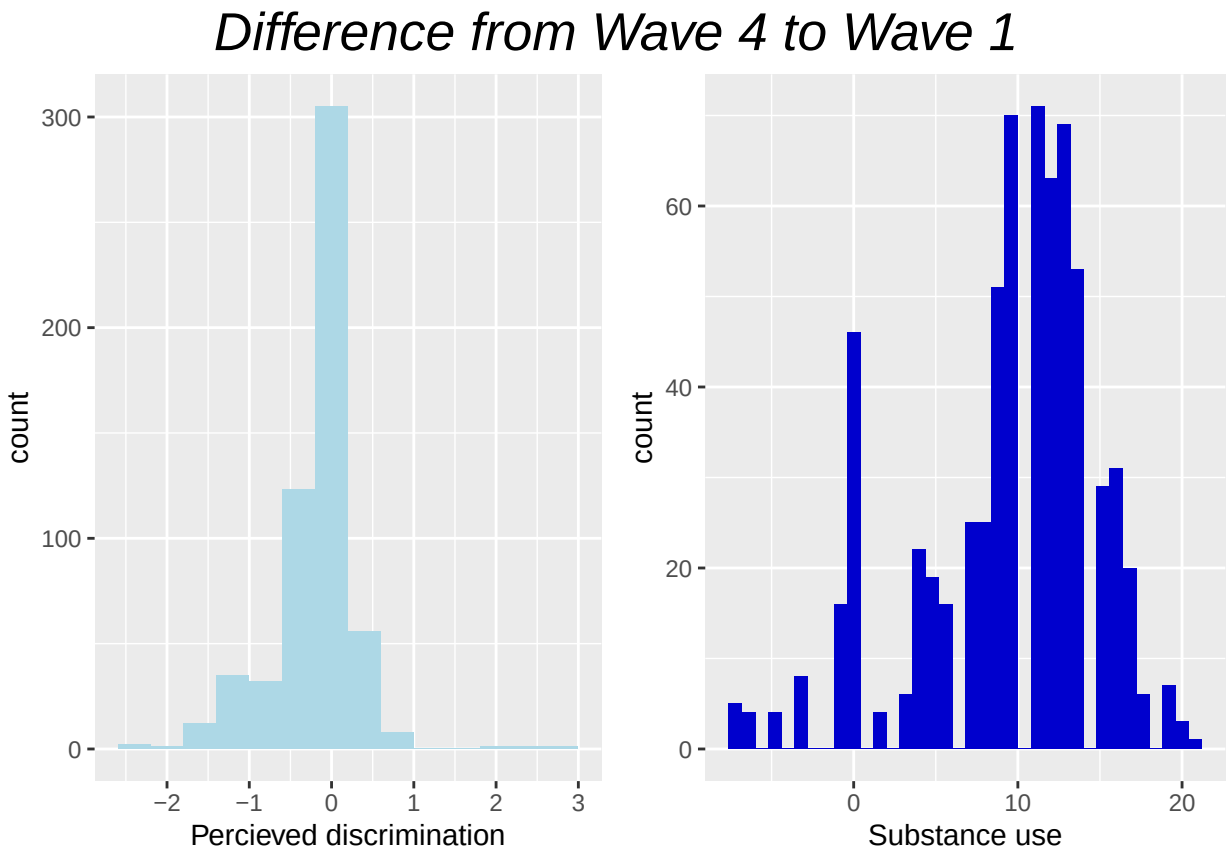
Conduct Analysis

1) Obtain difference scores and plot these new variables.

```
t1 <- ggplot(data = subset(w1234, !is.na(difference_dscr)), mapping = aes(difference_dscr))
t2 <- ggplot(data = subset(w1234, !is.na(difference_atod)), mapping = aes(difference_atod))

g1 <- t1 + geom_histogram(fill = "lightblue", binwidth = 0.4) +
    labs(x = "Percieved discrimination")
g2 <- t2 + geom_histogram(fill = "mediumblue", binwidth = 0.8) +
    labs(x = "Substance use")

grid.arrange(g1, g2, ncol = 2, nrow = 1, top = textGrob("Difference from Wave 4 to Wave 1",
    gp = gpar(fontsize = 20, font = 3)))
```



Percieved discrimination appears to stay relatively constant from wave 1 to wave 4. It has a smaller distrubtion. Percieved discrimination slightly trends towards a decrease in percieved discrmination. Substance use has a wider distribution. From wave 4 to wave 1, substance use increases.

2) Find and report the median for each of your difference score variables

```
median_diff_dscr <- median(w1234$difference_dscr, na.rm = TRUE)
median_diff_dscr
```

Median of percieved discrimination difference from wave 4 to wave 1

```
## [1] 0
```

```
median_diff_atod <- median(w1234$difference_atod, na.rm = TRUE)
median_diff_atod
```

Median of substance use difference from wave 4 to wave 1

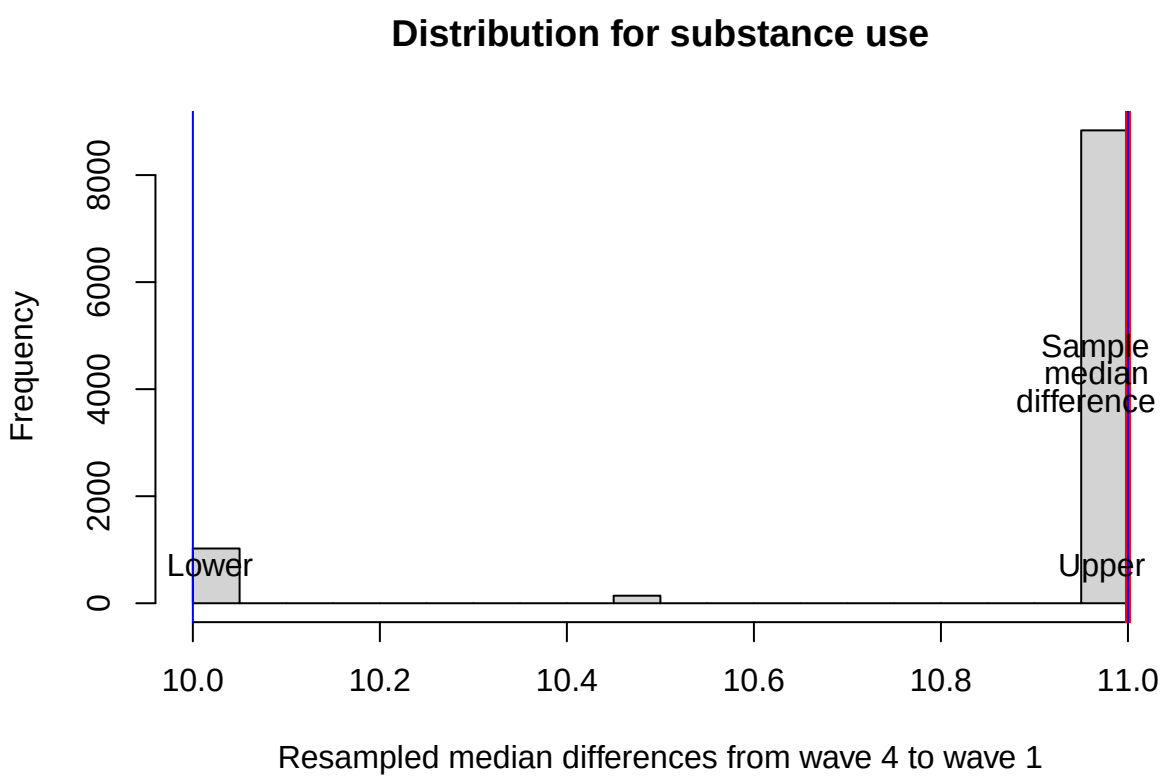
```
## [1] 11
```

3) Bootstrap to obtain confidence intervals at the 95% level

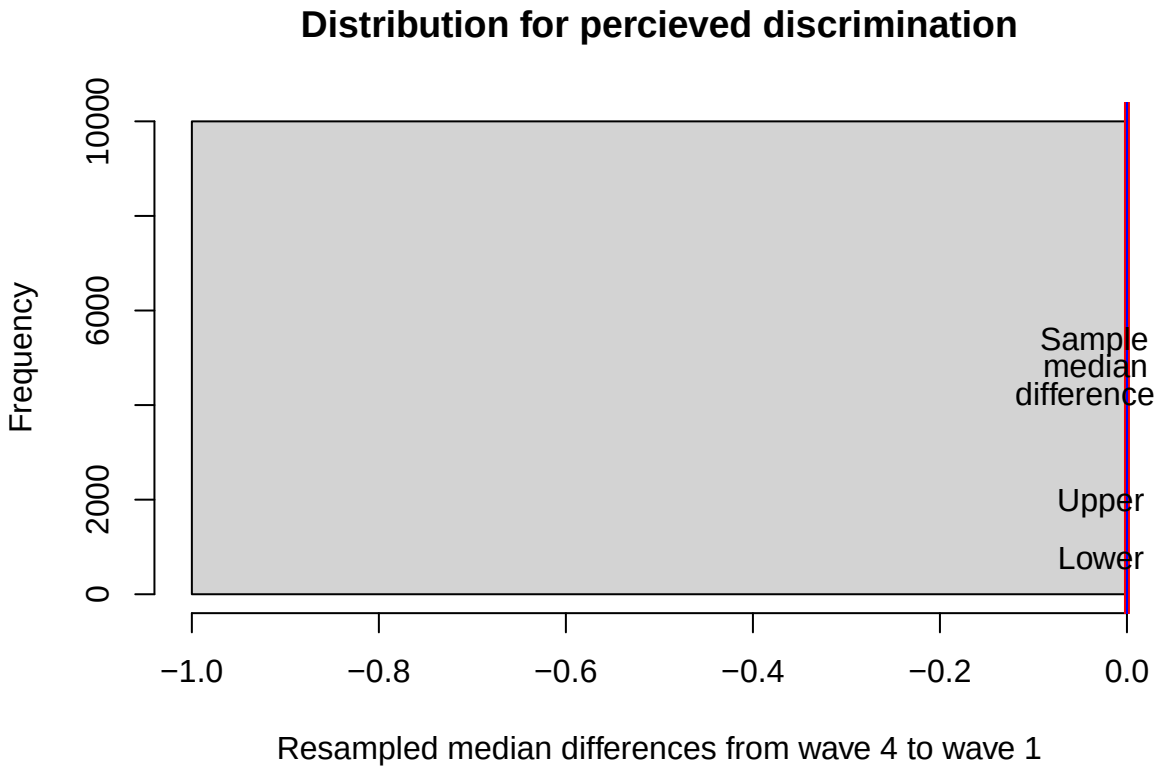
```
atodMedians <- matrix(, 10000, 1)
dscrMedians <- matrix(, 10000, 1)

for (reps in 1:10000) {
    newSample <- sample(w1234$difference_atod, replace = TRUE)
    atodMedians[reps] <- median(newSample, na.rm = TRUE)
    newSample <- sample(w1234$difference_dscr, replace = FALSE)
    dscrMedians[reps] <- median(newSample, na.rm = TRUE)
}
```

```
hist(atodMedians, xlab = "Resampled median differences from wave 4 to wave 1",
    main = paste("Distribution for substance use"))
abline(v = median_diff_atod, col = "red", lwd = 3)
abline(v = quantile(atodMedians, c(0.025, 0.975))[1], col = "blue")
abline(v = quantile(atodMedians, c(0.025, 0.975))[2], col = "blue")
text(quantile(atodMedians, c(0.025, 0.975))[2], 300, "Upper",
    adj = c(0.8, -0.5))
text(quantile(atodMedians, c(0.025, 0.975))[1], 300, "Lower",
    adj = c(0.3, -0.5))
text(median_diff_atod, 310, "Sample", adj = c(0.8, -10))
text(median_diff_atod, 310, "median", adj = c(0.8, -8.8))
text(median_diff_atod, 310, "difference", adj = c(0.8, -7.6))
```




```
hist(dscrMedians, xlab = "Resampled median differences from wave 4 to wave 1",
    main = paste("Distribution for percieved discrimination"))
abline(v = median_diff_dscr, col = "red", lwd = 3)
abline(v = quantile(dscrMedians, c(0.025, 0.975))[1], col = "blue")
abline(v = quantile(dscrMedians, c(0.025, 0.975))[2], col = "blue")
text(quantile(dscrMedians, c(0.025, 0.975))[2], 300, "Upper",
    adj = c(0.8, -3))
text(quantile(dscrMedians, c(0.025, 0.975))[1], 300, "Lower",
    adj = c(0.8, -0.5))
text(median_diff_dscr, 310, "Sample", adj = c(0.8, -10))
text(median_diff_dscr, 310, "median", adj = c(0.8, -8.8))
text(median_diff_dscr, 310, "difference", adj = c(0.8, -7.6))
```



4) Make an inference

For differences across wave 4 and wave 1, substance use is significantly different from zero. This is demonstrated by the 95% confidence interval ranging from 10 to 11, which is far from zero. Perceived discrimination is not significantly difference from zero. This is illuminated by the histogram containing only zero, which means that no other median out of 10000 re-samples was found to be anything other than zero.