Lab Assignment 0 BIS 505b

Wenxin Xu

2/14/2021

Instructions

Your Lab Assignment is in this **R** Markdown document and covers the material from **Lab 0** (2/1/21) and **Lab 1** (2/8/21). Use **R** to answer all questions. Perform your analyses in different code chunks within this document and provide interpretations and responses in the **R** Markdown text. Your **R** Markdown document should always display your **R** code along with its output or results used to support your answers. The three code chunks at the beginning of this document do not need to be displayed (hence the code chunk option `include=FALSE`). Weave together narrative text with output to create a logical flow in your write-up. Unless stated in the question, you should never include **just** code/output as your response to a question. You should always include some text/comment/interpretation in natural language, even it's something as simple as stating your finding in a complete sentence, e.g., "The average income in those over 30 in this sample is $32,148.33." You'll find that inline **R** coding is nice feature of **R** Markdown.

To submit your assignment, **(1)** compile (knit) your final .Rmd file to .html, **(2)** open the .html file in a web browser, **(3)** print to .pdf, and **(4)** upload your PDF file to Canvas before the assignment due date.

You may keep the sections on **Public Health Application**, **Data Background** and **Data Key** in your submission if you wish. Perform your work in the **Assignment** section below. The goals of this assignment are to become comfortable using **R** Markdown and to begin describing the `hgb` data.

Public Health Application

One of the most widely used pesticides in the United States is **atrazine**, a triazine herbicide. Atrazine's main use is to control broadleaf and grassy weeds with the most common sites of application being corn, sugarcane, and sorghum. Once introduced into the environment, atrazine is not easily broken down and has been shown to persist for long periods. This persistence provides ample opportunity for water system contamination, including drinking water. A large number of animal studies have been conducted that demonstrate mixed results concerning the link between atrazine and several adverse health effects including reproductive outcomes and cancer.

Data Background

We will use data from a study conducted to investigate the impacts of herbicide exposure on maternal health. For this study, 995 pregnant women from a large farming community were recruited during their initial prenatal visits. The groundwater in the region is known to be exposed to herbicides. At the initial visit (approximately week 9 of pregnancy), the women were surveyed about their planned water drinking habits for the duration of the pregnancy. In particular, women were asked whether they plan to drink water, “Only from the tap,” “Only from bottled water,” or “From both the tap and bottle” [`group`]. For the purposes of this study, filtered water was grouped with bottled water. Of the 995 pregnant women, 275 planned to drink only from the tap, 320 planned to drink only bottled water, and 400 planned to drink from both sources. At the initial visit, the investigators also collected information on the number of previous births [`parity`], pre-pregnancy smoking status [`psmoke`] and weight [`wt0`], income [`income`], and years of schooling [`edyrs`]. Weight (lb) at the end [`wt1`] of the pregnancy was also recorded. Hemoglobin measurements were taken from these women at the initial visit at week 9 [`hgb9`] and throughout their pregnancy: weeks 12, 24, and 36 [`hgb36`].

Only women who had resided in the region for the last 10 years, had singleton births after week 36 (full-term), and were compliant with their water consumption plans were included in the analysis, resulting in a final sample of 270 who drank only from the tap, 315 who drank only bottled water, and 394 who drank from both sources. Women in the tap-water-only or the both-sources group were also asked to keep records of the amount of tap water consumed throughout the pregnancy. For each woman, the amount of tap water consumed (L) over the course of the pregnancy was recorded [`water`].

A CSV file [`hgb.csv`] is provided which contains data from the women in the study. The outcome of interest is the hemoglobin change from week 9 to week 36, or `hgb36-hgb8` . Hemoglobin is known to generally decrease during pregnancy, and this is reflected by a negative calculated change in hemoglobin. Larger negative values of change correspond to greater decreases in hemoglobin during pregnancy. The research question is interested in determining if there is a difference in the hemoglobin change in women who are exposed to herbicides versus those who are not.

Data Key – `hgb.csv`

Variable Name	Definition
`id`	Unique identifier for each subject
`group`	Water consumption group
	1 = Tap Water
	2 = Bottled/Filtered Water
	3 = Tap/Bottled/Filtered
`age`	Age at initial visit
`edyrs`	Years of schooling
`income`	Annual household income (ten-thousand dollars)
`wt0`	Pre-pregnancy weight (lb)

Variable Name	Definition
`wt1`	Weight at end of pregnancy (lb)
`parity`	Number of previous births
	0 = None
	1 = One
	2 = Two
	3 = Three or more
`prenatal`	Adequate prenatal care
	0 = No
	1 = Yes
`psmoke`	Pre-pregnancy smoker
	0 = No
	1 = Yes
`hgb9`	Week 9 hemoglobin (g/dL)
`hgb36`	Week 36 hemoglobin (g/dL)
`water`	Amount of tap water consumed (liters, L)

Assignment

1. Import the CSV file `hgb.csv` in the third code chunk above. Name your data frame `hgb`. [*Note*: No response is required for this question]

2. Answer the questions below:

a. [5 points] Use **R** to determine how many observations were imported and how many variables were imported.

```
dim(hgb)
```

```
## [1] 979  13
```

Answer: 979 observations and 13 were imported.

b. [5 points] Examine the structure of your data frame using the `str()` function. Were all variables imported as numerical (i.e., either numbers or integers)?

```
str(hgb)
```

```
## 'data.frame':    979 obs. of  13 variables:
##  $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ group   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age     : int  26 23 24 26 28 25 31 22 21 25 ...
##  $ edyrs   : int  17 13 19 13 18 13 15 14 20 19 ...
##  $ income  : num  2.18 2.21 2.21 2.21 2.23 ...
##  $ wt0     : num  154 158 132 168 186 ...
##  $ wt1     : num  196 200 165 206 226 ...
##  $ parity  : int  1 0 1 2 3 1 2 2 1 3 ...
##  $ prenatal: int  0 0 1 1 0 0 1 1 1 1 ...
##  $ psmoke  : int  1 0 1 0 0 0 0 0 1 0 ...
##  $ hgb9    : num  9.74 10 10.35 11.28 10.68 ...
##  $ hgb36   : num  6.89 7.7 7.51 8.43 7.49 ...
##  $ water   : num  299 267 308 293 302 ...
```

Answer: Not all variables imported as numerical, variables named `id`, `group`, `age`, `edyrs`, `parity`, `prenatal` and `psmoke` are integers.

3. Time for some data management.

a. [5 points] Add the following two variables to your `hgb` data frame by performing the necessary calculations. Use **R** to determine how many variables `hgb` now contains.

Variable Name	Definition
`wtgain`	Weight gain during pregnancy (lb) (positive if weight gained)
`change`	Change in hemoglobin between Week 9 and Week 36 (g/dL) (negative if hemoglobin decreases)

```
hgb$wtgain <- hgb$wt1 - hgb$wt0

hgb$change <- hgb$hgb36 - hgb$hgb9
```

Answer: Now `hgb` contains 15 variables.

b. [5 points] Print the first 10 observations of the variables `hgb9`, `hgb36` and `change` in your `hgb` data frame. Based on these observations, are you observing an increase or decrease in hemoglobin during pregnancy?

```
hgb[1:10, c("hgb9", "hgb36", "change")]
```

```
##     hgb9 hgb36 change
## 1   9.74  6.89  -2.85
## 2  10.00  7.70  -2.30
## 3  10.35  7.51  -2.84
## 4  11.28  8.43  -2.85
## 5  10.68  7.49  -3.19
## 6  13.16 10.08  -3.08
## 7  10.50  7.29  -3.21
## 8  11.51  8.29  -3.22
## 9  10.28  6.64  -3.64
## 10 10.43  7.19  -3.24
```

Answer: Based on these 10 observations, there was an decrease in hemoglobin during pregnancy.

c. [10 points] Sort your data frame by `change` from smallest to largest and save this sorted data frame as `hgbsort`. Because `change` is a negative value, in order to ultimately print the 10 largest changes (largest decreases) in `hgb`, we need to sort from smallest to largest instead of from largest to smallest. Print the first 10 observations from women in Group 2 using your data frame sorted on `change`. What do you notice about the variable `water` in these women? Does this seem correct?

```
hgbsort <- hgb[order(hgb$change),]

head(hgbsort[hgbsort$group == 2,])[1:10,]
```

```
##       id group age edyrs income   wt0   wt1 parity prenatal psmoke  hgb
9 hgb36
## 326  326     2  27     9  2.559 141.0 181.4      1        0      1 12.2
5  8.80
## 370  370     2  28    12  2.833 175.7 223.0      3        0      0 10.4
1  7.09
## 331  331     2  28    10  2.585 174.1 208.0      2        0      1 12.0
6  8.81
## 336  336     2  28    10  2.596 140.6 186.8      3        1      0 11.9
5  8.83
## 302  302     2  29     8  2.375 164.5 208.0      1        1      1 12.0
7  8.97
## 343  343     2  25     9  2.616 127.9 162.4      1        1      1 10.9
1  7.83
## NA    NA    NA  NA    NA     NA    NA    NA     NA       NA     NA    N
A    NA
## NA.1  NA    NA  NA    NA     NA    NA    NA     NA       NA     NA    N
A    NA
## NA.2  NA    NA  NA    NA     NA    NA    NA     NA       NA     NA    N
A    NA
## NA.3  NA    NA  NA    NA     NA    NA    NA     NA       NA     NA    N
A    NA
##       water wtgain change
## 326       0   40.4  -3.45
## 370       0   47.3  -3.32
## 331       0   33.9  -3.25
## 336       0   46.2  -3.12
## 302       0   43.5  -3.10
## 343       0   34.5  -3.08
## NA       NA     NA     NA
## NA.1     NA     NA     NA
## NA.2     NA     NA     NA
## NA.3     NA     NA     NA
```

Answer: The value of variable `water` in these women is 0. It seems correct because these women were in group 2 which only drank bottled water or filtered water, they didn't drink any tap water while variable `water` means the amount of tap water consumed.

d. [5 points] Based on the data key, do you see any categorical variables? If so, what are their variable names? What type of categorical variable is each one (dichotomous, nominal, or ordinal)?

```
str(hgb)
```

```
## 'data.frame':    979 obs. of  15 variables:
##  $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ group   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age     : int  26 23 24 26 28 25 31 22 21 25 ...
##  $ edyrs   : int  17 13 19 13 18 13 15 14 20 19 ...
##  $ income  : num  2.18 2.21 2.21 2.21 2.23 ...
##  $ wt0     : num  154 158 132 168 186 ...
##  $ wt1     : num  196 200 165 206 226 ...
##  $ parity  : int  1 0 1 2 3 1 2 2 1 3 ...
##  $ prenatal: int  0 0 1 1 0 0 1 1 1 1 ...
##  $ psmoke  : int  1 0 1 0 0 0 0 0 1 0 ...
##  $ hgb9    : num  9.74 10 10.35 11.28 10.68 ...
##  $ hgb36   : num  6.89 7.7 7.51 8.43 7.49 ...
##  $ water   : num  299 267 308 293 302 ...
##  $ wtgain  : num  42.3 41.5 33 38.9 39.2 37.1 53.4 32.3 30 40.1 ...
##  $ change  : num  -2.85 -2.3 -2.84 -2.85 -3.19 -3.08 -3.21 -3.22 -3.64
-3.24 ...
```

Answer: The categorical variables are `id` (type: nominal), `group` (type: nominal), `parity` (type: ordinal), `prenatal` (type: dichotomous) and `psmoke` (type: dichotomous).

e. [5 points] Create factor or ordinal variable versions of these existing categorical variables in the `hgb` data frame. Refer to the **Data Key** above when labeling the levels of the factor variables. Please use the variable naming convention that we applied in **Lab 1** to the factor versions of these variables (i.e., `variablename_factor`). Use **R** to determine how many variables `hgb` contains at this stage.

```
hgb <- dplyr::mutate(hgb,
                     group_factor=factor(group,
                                         levels=c(1,2,3),
                                         labels=c("Tap Water","Bottled/Fil
tered Water","Tap/Bottled/Filtered")),

                     parity_factor=factor(parity,
                                          levels=c(0,1,2,3),
                                          labels=c("None","One","Two","Thr
ee or more"),
                                          ordered=TRUE),

                     prenatal_factor=factor(prenatal,
                                            levels=c(0,1),
                                            labels=c("No","Yes")),

                     psmoke_factor=factor(psmoke,
                                          levels=c(0,1),
                                            labels=c("No","Yes")))
```

Answer: Now there are 19 variables in the `hgb`.

f. [10 points] Create the following two new categorical variables in your `hgb` data frame. Be sure to also create a factor version of each. Use **R** to determine how many variables `hgb` now contains. Make sure both of your categorical variables (or their factor versions) are created correctly by reporting summary statistics (`min` and `max`, in particular) within each category. For

example, summarize `edyrs` for each level of `ed` or `ed_factor` and comment on what you would expect to see based on how the variable was created and if this expectation is confirmed.

Variable Name	Definition
`ed`	Educational attainment
	0 = Less than HS (years of schooling < 12)
	1 = HS/GED (years of schooling = 12)
	2 = Some college or more (years of schooling > 12)
`anemic9`	Presence of anemia at week 9
	0 = Not anemic (week 9 hemoglobin >= 11 g/dL)
	1 = Anemic (week 9 hemoglobin < 11 g/dL)

```
hgb$ed[hgb$edyrs < 12] <- 0

hgb$ed[hgb$edyrs == 12] <- 1

hgb$ed[hgb$edyrs > 12] <- 2

hgb$anemic9[hgb$hgb9 >= 11] <- 0

hgb$anemic9[hgb$hgb9 < 11] <- 1

hgb <- dplyr::mutate(hgb,
                     ed_factor=factor(ed,
                                      levels=c(0,1,2),
                                      labels=c("Less than HS","HS/GHD","So
me college or more"),
                                      ordered=TRUE),

                     anemic9_factor=factor(anemic9,
                                           levels=c(0,1),
                                           labels=c("Not anemic","Anemic"
)))
```

Answer: Now `hgb` has 23 variables

```
summary(hgb$ed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.7344  1.0000  2.0000
```

```
hgb %>%
  group_by(ed_factor) %>%
  summarise(min(edyrs), max(edyrs))
```

```
## # A tibble: 3 x 3
##   ed_factor            `min(edyrs)` `max(edyrs)`
##   <ord>                       <int>        <int>
## 1 Less than HS                    8           11
## 2 HS/GHD                         12           12
## 3 Some college or more           13           20
```

Answer: The summary statistics of `edyrs` for each level of `ed` should be: max of level 0 < min of level 1 = 12 = max of level 1 < min of level 2; and the expectation is confirmed.

```
hgb %>%
  group_by(anemic9_factor) %>%
  summarise(min(hgb9), max(hgb9))
```

```
## # A tibble: 2 x 3
##   anemic9_factor `min(hgb9)` `max(hgb9)`
##   <fct>                <dbl>       <dbl>
## 1 Not anemic           11           13.3
## 2 Anemic                8.65        11.0
```

Answer: The summary statistics of `hgb9` for each level of `anemic9` should be: max of level 1 < 11 <= min of level 0; and the expectation is confirmed.

4. Next, we will create a subset of `hgb`.

a. [5 points] Create a subset called `hgb12` that only includes participants in the "pure" exposure groups (i.e., the tap water only group and the bottled/filtered water only group). After you create `hgb12`, run the function, `hgb12 <- droplevels(hgb12)`, which will drop any unused levels from factors in the data frame (this will be useful later). How many observations are in this new data frame?

```
hgb12 <- subset(hgb, group == 1 | group == 2)

hgb12 <- droplevels(hgb12) # drop any unused levels from factors in the da
ta frame
```

Answer: 585 observations are in the new data frame.

b. [5 points] The `table()` function can be used to report the number of observations (participants) in each group. After you create `hgb12`, run `table(hgb$group)` and `table(hgb12$group)` in a code chunk. Was your subset created correctly (i.e., does your subset only include subjects in the specified groups?)?

```
table(hgb$group)
```

```
##
##   1   2   3
## 270 315 394
```

```
table(hgb12$group)
```

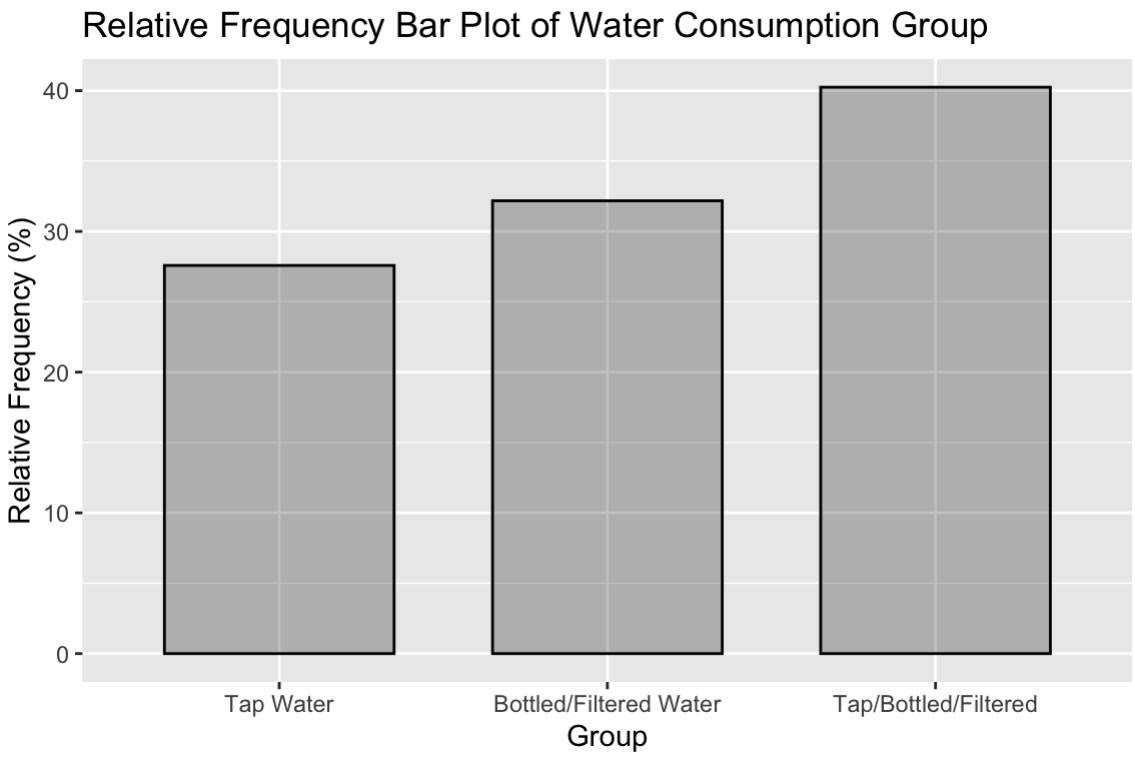
```
##
##   1   2
## 270 315
```

Answer: Yes, because the number of observations in group 1 and the number of observations in group 2 of `hgb12` are the same as those of `hgb`.

5. [10 points] Create the appropriate graph (choosing from either (1) a bar chart or (2) a histogram) to describe the distribution of women in each water consumption group. This graph should report the relative frequency (percentage) of women in the three water consumption groups. Include either a caption or a figure title. Which group contains the most women? Report the count and percentage (frequency and relative frequency) of women in this group.

```
library(ggplot2)

ggplot(data = hgb, aes(x = group_factor,
                       y = 100*(stat(count))/sum(stat(count)))) +
  geom_bar(col = "black", width = 0.7, alpha = 0.35) +
  labs(title = "Relative Frequency Bar Plot of Water Consumption Group",
       x = "Group", y = "Relative Frequency (%)")
```




```
100*table(hgb$group)/sum(table(hgb$group))
```

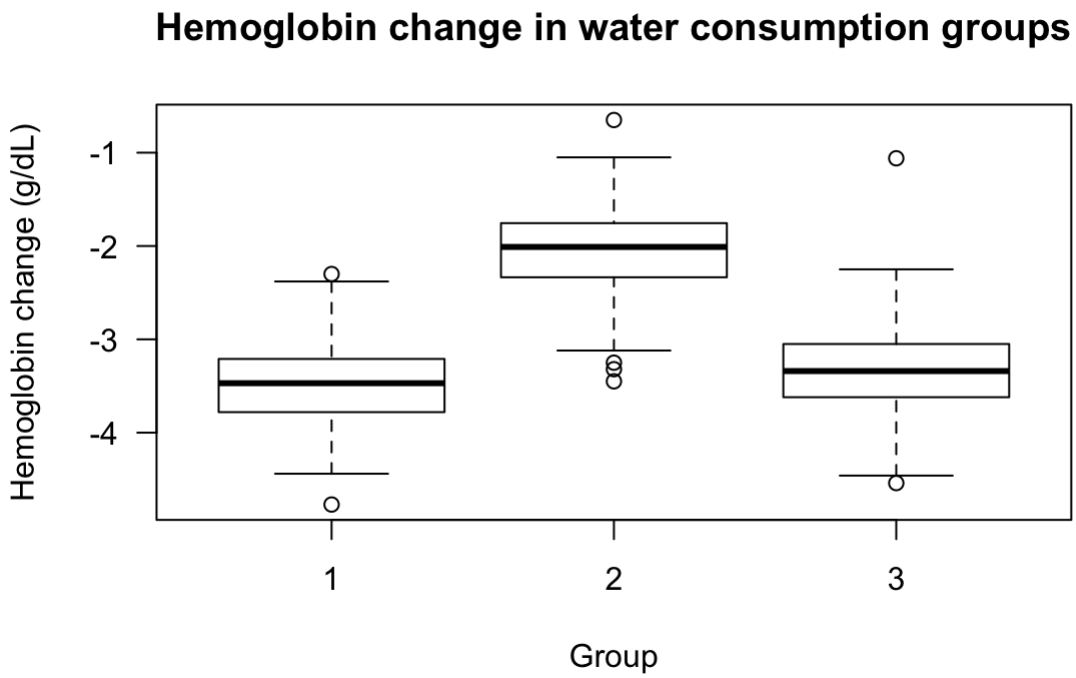
```
##
##        1        2        3
## 27.57916 32.17569 40.24515
```

Answer: There were 270 women in group1, accounting for 27.58%; 315 women in group2, accounting for 32.18%; 394 women in group3, accounting for 40.25%.

6. Let's explore baseline hemoglobin and hemoglobin change during pregnancy.

a. [10 points] Create boxplots of hemoglobin change in the three water consumption groups. Based on your visual inspection, how do the center and interquartile range in the combination tap/bottled/filtered water group compare to the other two groups? Does the range of the distribution (maximum, minimum shown in the boxplot) support the idea that the combination group potentially represents a more heterogeneous group of pregnant women? Do most of the women in the combination group have changes in hemoglobin that are similar to one of the other two groups?

```
boxplot(change ~ group, data=hgb,
        main="Hemoglobin change in water consumption groups",
        xlab="Group",
        ylab="Hemoglobin change (g/dL)",
        las=1)
```

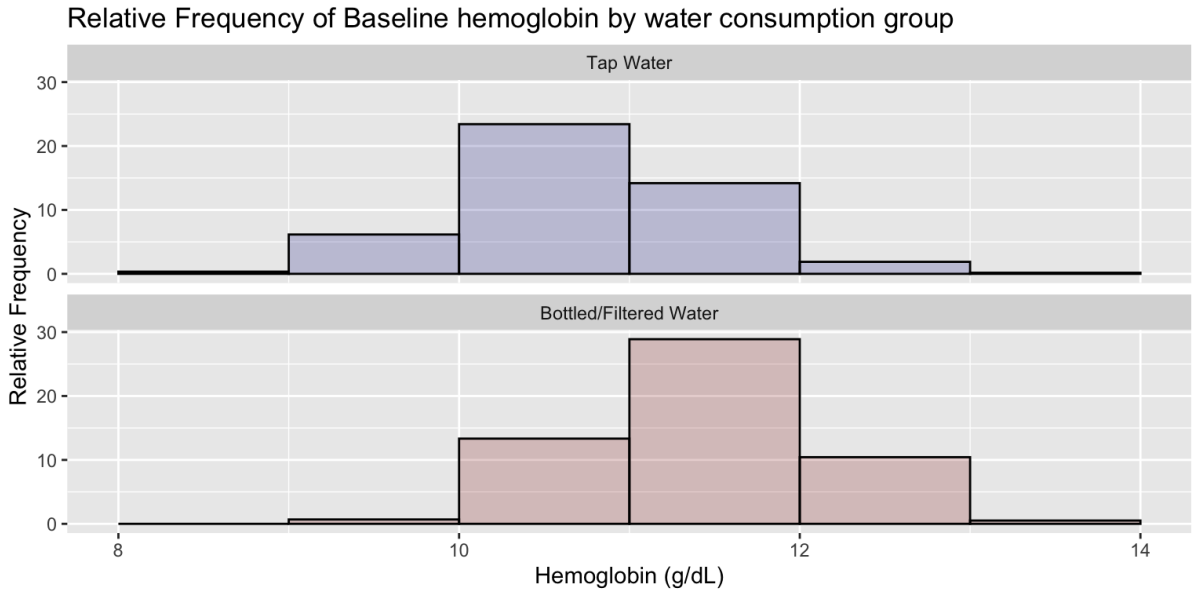


Answer: 1) The center and interquartile range in group3 are similar to group1, while are larger(absolute value of hemoglobin change) than group2. 2) Yes, the range of group3 is larger than both group1 and group2, so it support the idea that group3 is a more heterogeneous group. 3) Yes, it is similar to group1.

b. [10 points] Using `hgb12` , create histograms of baseline hemoglobin separately for the two water consumption groups (tap only, bottled/filtered only). Do the same for hemoglobin change. (Note: You are creating four total histograms in this question). The vertical axis of all histograms should display relative frequency. Choose a bin size that you think is appropriate. Does one group tend to have lower initial hemoglobin values or do they seem similar? Does one group tend to have larger changes in hemoglobin or do they seem similar? Describe what you see.

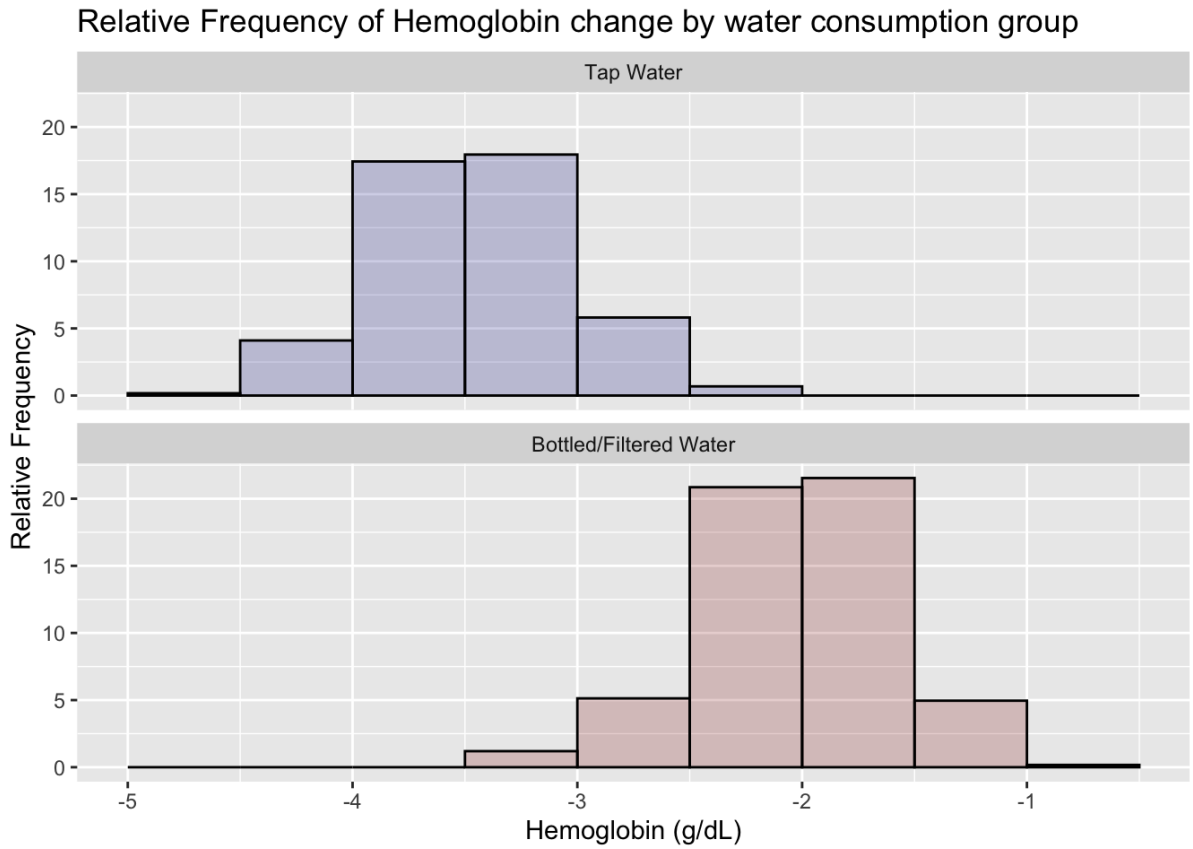
```
h1 <- ggplot(hgb12, aes(x = hgb9,  y = 100*stat(count) / sum(count),fill =
group_factor)) +
  geom_histogram(breaks=seq(from=8, to=14, by=1),
                 col = "black",
                 alpha = 0.2,
                 closed = "left") +
  labs(title = "Relative Frequency of Baseline hemoglobin by water consump
tion group",
       x = "Hemoglobin (g/dL)", y = "Relative Frequency", fill = "Group",c
ex.lab = 0.3) +
  scale_fill_manual(values = c("darkblue", "darkred")) +
  theme(legend.position = "none")

h1 + facet_wrap(~ group_factor, ncol = 1)
```




```
h2 <- ggplot(hgb12, aes(x = change,  y = 100*stat(count) / sum(count),fill
= group_factor)) +
  geom_histogram(breaks=seq(from=-5, to=-0.5, by=0.5),
                 col = "black",
                 alpha = 0.2,
                 closed = "left") +
  labs(title = "Relative Frequency of Hemoglobin change by water consumpti
on group",
       x = "Hemoglobin (g/dL)", y = "Relative Frequency", fill = "Group",c
ex.lab = 0.4) +
  scale_fill_manual(values = c("darkblue", "darkred")) +
  theme(legend.position = "none")

h2 + facet_wrap(~ group_factor, ncol = 1)
```



Answer: 1) Group1 tend to have lower initial hemoglobin values then group2. 2) Group1 tend to have larger changes in hemoglobin then group2.

c. [10 points] Using `hgb12`, report the mean and median baseline hemoglobin and hemoglobin change separately for the two water consumption groups (tap only, bottled/filtered only). Based on the histograms of baseline hemoglobin and hemoglobin change that you created in Question **6b**, would you expect the mean values of these variables to be approximately equal to, greater than, or less than their median values in each group? Why? Do your summary statistics support your expectations?

Because the histograms are about symmetric, we would expect the mean to be approximately equal to median .

```
hgb12 %>%
  group_by(group) %>%
  summarise(mean(hgb9), median(hgb9), mean(change), median(change))
```

```
## # A tibble: 2 x 5
##   group `mean(hgb9)` `median(hgb9)` `mean(change)` `median(change)`
##   <int>        <dbl>          <dbl>          <dbl>             <dbl>
## 1     1         10.7           10.8          -3.47             -3.47
## 2     2         11.4           11.4          -2.03             -2.01
```

Answer: I expect the mean value of baseline hemoglobin in group1 would be less than group2 because in the histogram 1 and 2 of Q6b, I found that group1 tend to have lower initial hemoglobin. I expect the mean value of hemoglobin change in group1 would be less than group2 because in the histogram 3 and 4 of Q6b, I found that group1 tend to have larger changes in hemoglobin.