Lab Assignment 2 BIS 505b

Wenxin Xu

3/14/2021

- Instructions
- Assignment

Instructions

This Lab Assignment uses the data from the study conducted to investigate the impacts of herbicide exposure on maternal health described in **Lab Assignment 0**, `hgb.csv` . Report any p-values that are less than 0.001 as <0.001 and round values reported in your narrative text to 3 decimal places.

Assignment

1. [5 points] Import the CSV file `hgb.csv` in the third code chunk above. Name your data frame `hgb` and re-create the variables `change` and `group_factor` that you created in **Lab Assignment 0**. [Note: No written response is required for this question. Display the code chunk(s) that perform the requested data management steps for this question.]

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

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

2. The **research question** is: Do differences exist in the *change in hemoglobin* in women who were exposed (tap water only, `group` = 1), marginally exposed (tap and bottled/filtered water, `group` = 3), and not exposed (bottled/filtered water only, `group` = 2) to herbicides in their drinking water?

a. [10 points] Complete **Table 1a** below (xxx values) with the following descriptive statistics: number of women in each water consumption group, number of non-missing values, mean (standard deviation), and median (range) of the *change in hemoglobin* in women who were exposed, marginally exposed, and not exposed to herbicides in water. Report any decimals to 2 decimal places in the table. [Note: You do not have to summarize your results in complete sentences.] Comment on what you observe in mean change in hemoglobin in the three groups.

```
summze <- function(x) c(n = length(x),
                        nonmissing = sum(!is.na(x)),
                        mean = round(mean(x, na.rm = TRUE), 2),
                        sdev = round(sd(x, na.rm = TRUE), 2),
                        median = round(median(x, na.rm = TRUE),2),
                        range = round(range(x, na.rm = TRUE), 2))

summary = aggregate(x = list(change = hgb$change),
          by = list(water = hgb$group_factor),
          FUN = summze)

summary_df <- cbind(summary[[ncol(summary)]],summary[-ncol(summary)])

row.names(summary_df) <- summary_df$water

summary_df = summary_df[,1:(ncol(summary_df)-1)]
```

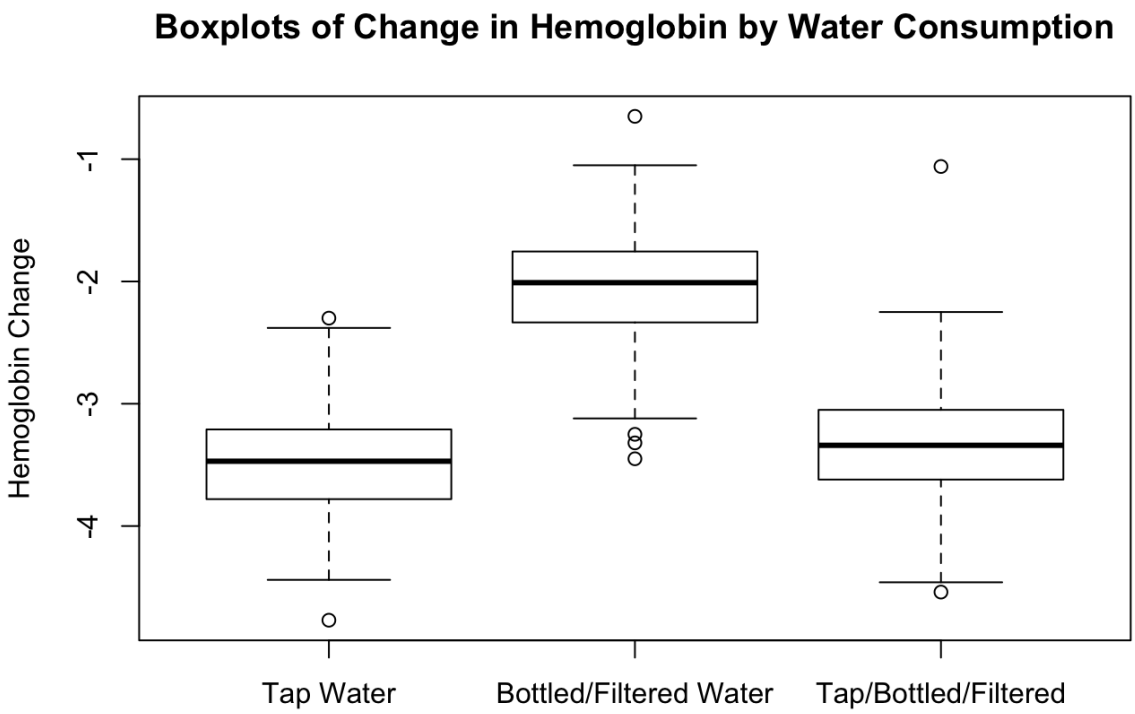
Table 1a. Characteristics of the Sample

Water Consumption Group	**Tap Only (N=270)**	**Bottled/Filtered Only (N=315)**	**Tap/Bottled/Filtered (N=394)**
Change in Hemoglobin			
N	270	315	394
Mean (SD)	-3.47 (0.42)	-2.03 (0.43)	-3.33 (0.42)
Median (Range)	-3.47 (-4.77, -2.3)	-2.01 (-3.45, -0.65)	-3.34 (-4.54, -1.06)

Comment: The difference of mean change in hemoglobin between exposed group 1 (or marginally exposed group 3) and not exposed group 2 is large while mean change of group 1 and group 3 is similar.

b. [5 points] Use boxplots to compare the distribution of *change in hemoglobin* in the three exposure groups. Comment on the similarity/differences between groups that you observe.

```
boxplot(change ~ group_factor, data = hgb,
        main = "Boxplots of Change in Hemoglobin by Water Consumption",
        xlab = "",
        ylab = "Hemoglobin Change")
```



Comment: Hemoglobin change in group 1 and group 3 is similar while Hemoglobin change in group 1 (or group 3) and group 2 is very different.

c. [15 points] Is there evidence against the hypothesis that the mean change in hemoglobin during pregnancy is equal in these three populations at the α = 0.05-level? **(i)** State the null and alternative hypotheses of this test. **(ii)** Using the "rule of thumb" presented in the lecture, does the constant variance assumption of analysis of variance seem justified? Explain. **(iii)** From your **R** output, report the value of the test statistic and p-value. **(iv)** State your statistical conclusion and your conclusion in the context of the problem.

(i) $H_0: \mu_1=\mu_2=\mu_3$ vs. $H_1: at\ least\ one\ \mu_i \neq \mu_j$

(ii)

```
maxsd <- max(summary_df$sdev)

minsd <- min(summary_df$sdev)

maxsd/minsd
```

```
## [1] 1.02381
```

Explanation: The constant variance assumption of analysis of variance is justified because the ratio of the largest group standard deviation (0.43) to the smallest group standard deviation (0.42) is equal to 1.024, which is **less than 2**.

(iii)

```
# ANOVA: Equal variance assumption
anova.change <- aov(change ~ group_factor, data = hgb)

res.change <- summary(anova.change)

res.change
```

```
##               Df Sum Sq Mean Sq F value              Pr(>F)
## group_factor   2  396.2  198.08    1097 <0.0000000000000002
## Residuals    976  176.2    0.18
```

The F test statistics is 1096.957. The p-value is 0.

(iv) Because p-value is less than 0.05, we have evidence to reject \(H_0\) and conclude the mean change in hemoglobin during pregnancy is not equal in all three populations at the \(\alpha\) = 0.05-level of significance.

d. [5 points] What is the estimate of the common within-group variance?

The estimate of common within-group variance \(s^2_W\) is 0.181.

e. [10 points] In question **2(c)**, if you found evidence that mean change in hemoglobin was not equal in all three groups of women, next determine which groups are significantly different from one another, maintaining an overall type I error rate of \(\alpha\) = 0.05. Report the Bonferroni-adjusted p-values for all pairwise comparisons. In which two groups do we observe the largest difference in mean hemoglobin change? Report and interpret this estimated difference in means between the two groups with the largest difference.

```
# Bonferroni adjusted p-values
res.pair = pairwise.t.test(hgb$change, hgb$group_factor, p.adjust.method =
"bonferroni")

res.pair
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  hgb$change and hgb$group_factor
##
##                          Tap Water            Bottled/Filtered Water
## Bottled/Filtered Water < 0.0000000000000002 -
## Tap/Bottled/Filtered   0.00021              < 0.0000000000000002
##
## P value adjustment method: bonferroni
```

The Bonferroni-adjusted p-values for tap water group and bottled/filtered water group is 0, for tap water group and tap/bottled/filtered water group is 0, and for bottled/filtered water group and tap/bottled/filtered water group is 0.

```
model.change <- lm(change ~ group_factor, data = hgb)
bontest <- lsmeans(model.change,
                   pairwise ~ group_factor,
                   adjust = "bonferroni")

res = summary(bontest$contrasts)

res
```

```
##  contrast                                           estimate     SE  df
t.ratio
##  Tap Water - (Bottled/Filtered Water)                 -1.436 0.0352 976
-40.758
##  Tap Water - (Tap/Bottled/Filtered)                   -0.134 0.0336 976
-3.998
##  (Bottled/Filtered Water) - (Tap/Bottled/Filtered)    1.302 0.0321 976
40.544
##  p.value
##  <.0001
##  0.0002
##  <.0001
##
## P value adjustment: bonferroni method for 3 tests
```

Conclusion: We observe largest difference in 2 pairs: exposed group vs not exposed group (adjusted p value = 0) with estimated difference in mean -1.436, and in marginally exposed group vs not exposed group (adjusted p value = 0), with estimated difference in mean -1.302.

f. [5 points] What is the critical value of the pairwise 2-sided Bonferroni t-test performed comparing the tap water only group (`group` = 1) and the bottled/filtered water only group (`group` = 2), assuming all three pairwise comparisons will be performed? Remember that the critical value of the Bonferroni tests is based on the \(\alpha^*\) level of significance in order to control the overall type I error at the \(\alpha = 0.05\) level. How does this critical value compare to the critical value of the same test performed at the \(\alpha\) level (instead of the \(\alpha^*\) level) (i.e., the unadjusted test)? Under which test is it more difficult to reject \(H_0\)? Why is this adjustment necessary?

```
# critical value of t test for group 1 and group 2
k = 3
c = k*(k-1)/2
alpha = 0.05
alpha_star = 0.05/c

# critical value of Bonferroni-adjusted t test
adjusted_critical = qt(1- (alpha_star/2), res$df[1])

# critical value of unadjusted t test
non_adjusted_critical = qt(1 - (alpha/2), res$df[1])
```

Answer: the critical value of Bonferroni-adjusted t test (2.398) is larger than that of unadjusted t test (1.962). Under Bonferroni-adjusted t test is more difficult to reject \(H_0\). Because perform all 3 pairwise t-tests will increase the chance of making a Type I error (falsely rejecting \(H_0\)), the familywise error rate is 0.143.

g. [10 points] Summarize your results and conclusions in a few sentences. Comment on whether your test results support what you saw summarized numerically in **2(a)** and graphically in **2(b)**. Our goal is to address the **research question** above.

Answer: There is significant difference between women who were exposed (tap water only, group = 1) and not exposed (bottled/filtered water only, group = 2) (adjusted p value = 0), and significant difference between women who were marginally exposed (tap and bottled/filtered water, group = 3) and not exposed (adjusted p value = 0). These findings are consistent with results in table and boxplot.

3. Remember that this is an observational study in that women are not assigned or randomized to their exposure group. The **research question** is: Do differences exist in the *baseline hemoglobin* `hgb9` in women who were exposed (tap water only), marginally exposed (tap and bottled/filtered water), and not exposed (bottled/filtered water only) to herbicides in their drinking water?

a. [10 points] Complete **Table 1b** below (xxx values) with the following descriptive statistics: number of non-missing values, mean (standard deviation), and median (range) of *baseline hemoglobin* in women who were exposed, marginally exposed, and not exposed to herbicides in water. Report any decimals to 2 decimal places in the table. [Note: You do not have to summarize your results in complete sentences.] Comment on what you observe in mean baseline hemoglobin in the three groups.

```
summary_2 = aggregate(x = list(hgb9 = hgb$hgb9),
          by = list(water = hgb$group_factor),
          FUN = summze)

summary_df_2 <- cbind(summary_2[[ncol(summary_2)]],summary_2[-ncol(summary
_2)])

row.names(summary_df_2) <- summary_df_2$water

summary_df_2 = summary_df_2[,1:(ncol(summary_df_2)-1)]
```

Table 1b. Characteristics of the Sample (continued)

Water Consumption Group	**Tap Only (N=270)**	**Bottled/Filtered Only (N=315)**	**Tap/Bottled/Filtered (N=394)**
Week 9 Hemoglobin			
N	270	315	394
Mean (SD)	10.73 (0.74)	11.42 (0.65)	11.29 (0.69)
Median (Range)	10.75 (8.65, 13.16)	11.41 (9.68, 13.28)	11.29 (9.05, 13.3)

Comment: mean baseline hemoglobin in exposed group 1 (10.73) is different from that in not exposed group 2 (11.42) and marginally exposed group 3 (11.29), while mean baseline hemoglobin in group 2 and group 3 is similar.

b. [15 points] Is there evidence against the hypothesis that the mean baseline hemoglobin is equal in these three populations at the α = 0.05-level? **(i)** State the null and alternative hypotheses of this test. **(ii)** Again check the "rule of thumb" to assess the ANOVA constant variance assumption. **(iii)** From your **R** output, report the value of the test statistic and p-value. **(iv)** State your statistical conclusion and your conclusion in the context of the problem.

(i)

$H_0: \mu_1=\mu_2=\mu_3$ vs. $H_1: at\ least\ one\ \mu_i \neq \mu_j$

(ii)

```
maxsd <- max(summary_df_2$sdev)

minsd <- min(summary_df_2$sdev)

maxsd/minsd
```

```
## [1] 1.138462
```

Explanation: The constant variance assumption of analysis of variance is justified because the ratio of the largest group standard deviation (0.74) to the smallest group standard deviation (0.65) is equal to 1.138, which is **less than 2**.

(iii)

```
# ANOVA: Equal variance assumption
anova.hgb9 <- aov(hgb9 ~ group_factor, data = hgb)

res.hgb9 <- summary(anova.hgb9)

res.hgb9
```

```
##               Df Sum Sq Mean Sq F value              Pr(>F)
## group_factor   2   78.5   39.26   81.97 <0.0000000000000002
## Residuals    976  467.5    0.48
```

The F test statistics is 81.97. The p-value is 0.

(iv) Because p-value is less than 0.05, we have evidence to reject H_0 and conclude the mean baseline hemoglobin is not equal in all three populations at the α = 0.05-level of significance.

c. [10 points] [Note: No analyses are required for question.] If some groups of women tend to begin the study with higher or lower levels of hemoglobin, do you think this could affect how much their hemoglobin can/will decrease during pregnancy? Since we want to study how exposure to herbicides (water consumption group is our proxy for herbicide exposure) affects

change in hemoglobin during pregnancy, do you think that differences in baseline hemoglobin should be something we account (“control”) for in our analysis of the relationship between change in hemoglobin and exposure group?

Answer: I think this could affect how much their hemoglobin can/will decrease during pregnancy. I think differences in baseline hemoglobin should be controled in our analysis of the relationship between change in hemoglobin and exposure group.