Lab Assignment 1 BIS 505b

Wenxin Xu

2/21/2021

- 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`. The analyses that you perform will use data from the tap water consumption group (`group` = 1) and the bottled water consumption group (`group` = 2). The women who consumed only tap water (`group` = 1) are considered to be "exposed" because they were exposed to the herbicides found in the region's groundwater.

When testing, use either an *unpooled t-test* (default in `t.test()`) or *continuity-corrected binomial test of proportions* (default in `prop.test()`)/*continuity-corrected chi-square test* (default in `chisq.test()`), as appropriate. 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. Begin with data management. [Note: No written responses are required for this question. Display the code chunks that perform the requested data management steps for this question.]

a. [5 points] Import the CSV file `hgb.csv` in the third code chunk above. Name your data frame `hgb` and re-create the variables `wtgain`, `change`, `ed`, `anemic9`, `group_factor`, `parity_factor`, `prenatal_factor`, `psmoke_factor`, `ed_factor`, and `anemic9_factor` that you created in **Lab Assignment 0**. After these steps, `hgb` should contain 23 variables.

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

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

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,
                     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(1,0),
                                            labels=c("Yes","No")),

                     psmoke_factor=factor(psmoke,
                                          levels=c(1,0),
                                            labels=c("Yes","No")),
                     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(1,0),
                                           labels=c("Anemic","Not anemic"
)))
```

```
ncol(hgb)
```

```
## [1] 23
```

Answer: Now `hgb` contains 23 variables.

b. [5 points] Create a subset called `hgb12` that only includes participants in the tap water only consumption group and the bottled/filtered water only consumption group. You will work with this data frame in this Lab Assignment. After you create `hgb12`, run the code, `hgb12 <- droplevels(hgb12)`.

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

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

2. We would like to determine if certain characteristics differ significantly between these two water consumption groups. Our main comparison of interest is the difference in hemoglobin change during pregnancy in the tap water only group and the bottled/filtered water only group.

a. [10 points] Determine if pregnant women who drink only tap water and pregnant women who drink only bottled water are different, on average, with respect to the primary outcome, hemoglobin change over the course of pregnancy. **(i)** State the null and alternative hypotheses of this test and **(ii)** report the name of the test that you performed. **(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. **(v)** If a significant difference is found, state which group has a larger average hemoglobin decline and report the average decline in each group.

```
tt = t.test(change ~ group_factor, data = hgb12, var.equal=FALSE, alternat
ive="two.sided")

tt
```

```
##
##  Welch Two Sample t-test
##
## data:  change by group_factor
## t = -40.46, df = 572.77, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.506126 -1.366668
## sample estimates:
##              mean in group Tap Water mean in group Bottled/Filtered Wat
er
##                             -3.469000                             -2.0326
03
```

i. $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$
ii. Perform a two-sample unpooled 2-sided t-test
iii. t statistic = -40.46; p-value = 0
iv. Reject $H_0 : \mu_1 = \mu_2$ and conclude $H_1 : \mu_1 \neq \mu_2$. That is, there is evidence to conclude that mean hemoglobin change is significantly different in pregnant women who drink only tap water and pregnant women who drink only bottled water (p-value from unpooled test 0 is less than significant level 0.05).
v. Tap water consumption group has a larger average hemoglobin decline (-3.469 g/dL) than bottled/filtered water consumption group (-2.033 g/dL).

b. [10 points] Report the point estimate of the difference in the mean hemoglobin change between the tap water consumption group and the bottled/filtered water consumption group and the 95% confidence interval for the difference in means. Does the result of your hypothesis test in **2(a)** seem reasonable given the confidence interval for the difference in means? Why?

The point estimate of the difference in the mean hemoglobin change between the tap water consumption group and the bottled/filtered water consumption group is -1.436 g/dL. The 95% unpooled confidence interval for the mean difference is (-1.506, -1.367). The result of my hypothesis test in 2(a) seem reasonable given the confidence interval, because this confidence interval does not contain 0, the value hypothesized for the difference in means under H_0 (i.e., $H_0 : \mu_1 - \mu_2 = 0$).

c. Determine if mean **(i)** age, **(ii)** income, **(iii)** weight gain, **(iv)** week 9 hemoglobin, and **(v)** final hemoglobin are significantly different in the two water consumption groups. Report the name of the test performed, state your statistical conclusion, p-value, and your conclusion in the context of the problem. If a significant difference is found, state which group has a higher or lower average value of the variable being compared, report a point estimate of the difference in means, and report the 95% confidence interval for the difference in means.

- **(i).** [5 points] Age

```
tt_age = t.test(age ~ group_factor, data = hgb12, var.equal=FALSE, alterna
tive="two.sided")

tt_age
```

```
##
##  Welch Two Sample t-test
##
## data:  age by group_factor
## t = -14.705, df = 497.15, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.389588 -1.826285
## sample estimates:
##              mean in group Tap Water mean in group Bottled/Filtered Wat
er
##                             25.02222                                27.130
16
```

Perform a two-sample unpooled 2-sided t-test.Reject $H_0 : \mu_1 = \mu_2$ and conclude $H_1 : \mu_1 \neq \mu_2$. That is, there is evidence to conclude that mean age is significantly different in pregnant women who drink only tap water and pregnant women who drink only bottled water (p-value from unpooled test 0 is less than significant level 0.05).Tap water consumption group has a smaller average age (25.022 years old) than bottled/filtered water consumption group (27.13 years old). The point estimate of the difference in the mean age is -2.108 years old. The 95% unpooled confidence interval for the mean difference is (-2.39, -1.826).

- **(ii).** [5 points] Income (multiply point estimate of difference in means and its and CI by 10,000 to report on the scale of dollars)

```
tt_income = t.test(income ~ group_factor, data = hgb12, var.equal=FALSE, a
lternative="two.sided")

tt_income
```

```
##
##  Welch Two Sample t-test
##
## data:  income by group_factor
## t = -2.299, df = 577.35, p-value = 0.02186
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.56756053 -0.04458833
## sample estimates:
##              mean in group Tap Water mean in group Bottled/Filtered Wat
er
##                             3.753519                             4.0595
93
```

Perform a two-sample unpooled 2-sided t-test.Reject $H_0 : \mu_1 = \mu_2$ and conclude $H_1 : \mu_1 \neq \mu_2$. That is, there is evidence to conclude that mean income is significantly different in pregnant women who drink only tap water and pregnant women who drink only bottled water (p-value from unpooled test 0.022 is less than significant level 0.05).Tap water consumption group has a lower average income (37535.185 dollars) than bottled/filtered water consumption group (40595.929 dollars). The point estimate of the difference in the mean income is -3060.744 dollars. The 95% unpooled confidence interval for the mean difference is (-5675.605, -445.883).

- **(iii).** [5 points] Weight Gain

```
tt_wtgain = t.test(wtgain ~ group_factor, data = hgb12, var.equal=FALSE, a
lternative="two.sided")

tt_wtgain
```

```
##
##  Welch Two Sample t-test
##
## data:  wtgain by group_factor
## t = -4.9861, df = 482.12, p-value = 0.0000008609
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.919220 -1.268822
## sample estimates:
##              mean in group Tap Water mean in group Bottled/Filtered Wat
er
##                             40.01074                             42.104
76
```

Perform a two-sample unpooled 2-sided t-test.Reject $H_0 : \mu_1 = \mu_2$ and conclude $H_1 : \mu_1 \neq \mu_2$. That is, there is evidence to conclude that mean weight gain is significantly different in pregnant women who drink only tap water and pregnant women who drink only

bottled water (p-value from unpooled test 0 is less than significant level 0.05).Tap water consumption group has a smaller average weight gain (40.011 lb) than bottled/filtered water consumption group (42.105 lb). The point estimate of the difference in the mean weight gain is -2.094 lb. The 95% unpooled confidence interval for the mean difference is (-2.919, -1.269).

- **(iv).** [5 points] Week 9 Hemoglobin

```
tt_hgb9 = t.test(hgb9 ~ group_factor, data = hgb12, var.equal=FALSE, alter
native="two.sided")

tt_hgb9
```

```
##
##  Welch Two Sample t-test
##
## data:  hgb9 by group_factor
## t = -11.991, df = 539.17, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8082642 -0.5807199
## sample estimates:
##              mean in group Tap Water mean in group Bottled/Filtered Wat
er
##                              10.72856                                11.423
05
```

Perform a two-sample unpooled 2-sided t-test.Reject $H_0 : \mu_1 = \mu_2$ and conclude $H_1 : \mu_1 \neq \mu_2$. That is, there is evidence to conclude that mean week 9 hemoglobin is significantly different in pregnant women who drink only tap water and pregnant women who drink only bottled water (p-value from unpooled test 0 is less than significant level 0.05).Tap water consumption group has a smaller average week 9 hemoglobin (10.729 g/dL) than bottled/filtered water consumption group (11.423 g/dL). The point estimate of the difference in the mean week 9 hemoglobin is -0.694 g/dL. The 95% unpooled confidence interval for the mean difference is (-0.808, -0.581).

- **(v).** [5 points] Final Hemoglobin

```
tt_hgb36 = t.test(hgb36 ~ group_factor, data = hgb12, var.equal=FALSE, alt
ernative="two.sided")

tt_hgb36
```

```
##
##  Welch Two Sample t-test
##
## data:  hgb36 by group_factor
## t = -31.728, df = 540.78, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.262818 -1.998960
## sample estimates:
##              mean in group Tap Water mean in group Bottled/Filtered Wat
er
##                             7.259556                                9.3904
44
```

Perform a two-sample unpooled 2-sided t-test.Reject $H_0 : \mu_1 = \mu_2$ and conclude $H_1 : \mu_1 \neq \mu_2$. That is, there is evidence to conclude that mean final hemoglobin is significantly different in pregnant women who drink only tap water and pregnant women who drink only bottled water (p-value from unpooled test 0 is less than significant level 0.05).Tap water consumption group has a smaller average final hemoglobin (7.26 g/dL) than bottled/filtered water consumption group (9.39 g/dL). The point estimate of the difference in the mean final hemoglobin is -2.131 g/dL. The 95% unpooled confidence interval for the mean difference is (-2.263, -1.999).

Total answer: Mean age, income, weight gain, week 9 hemoglobin and final hemoglobin are all significantly different in the two water consumption groups.

3. Another way of looking at hemoglobin is to dichotomize into an indicator of presence of anemia. You created this indicator in question **1(a)** using the week 9 hemoglobin value `anemic9_factor`.

a. [10 points] Determine if the prevalence of anemia (“disease”) at week 9 of pregnancy is significantly different in pregnant women who drink only tap water (“exposed”) and pregnant women who drink only bottled/filtered water (“unexposed”). **(i)** State the null and alternative hypotheses of this test and **(ii)** report the name of the test that you performed. **(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. **(v)** If a significant difference is found, state which group has a larger estimated prevalence of anemia at week 9 and report the proportion with anemia at week 9 in each group.

```
tab2p = table(hgb12$group_factor, hgb12$anemic9_factor,
              dnn = c("Water consumption", "Anemic"))
tab2p
```

```
##                          Anemic
## Water consumption        Anemic Not anemic
##   Tap Water                 175         95
##   Bottled/Filtered Water     82        233
```

```
pt = prop.test(tab2p,alternative = "two.sided")

pt
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tab2p
## X-squared = 87.211, df = 1, p-value < 0.00000000000000022
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.3096063 0.4660551
## sample estimates:
##    prop 1    prop 2
## 0.6481481 0.2603175
```

i. $H_0 : p_1 = p_2$ vs. $H_1 : p_1 \neq p_2$
ii. Perform a two-sample binomial test of proportions
iii. Chi statistic = 87.211; p-value = 0 *iii. when the p-value is very small, it is better to say p-value< 0.001*
iv. Reject $H_0 : p_1 = p_2$ and conclude $H_1 : p_1 \neq p_2$. That is, there is evidence to conclude that the prevalence of anemia at week 9 of pregnancy is significantly different in pregnant women who drink only tap water ("exposed") and pregnant women who drink only bottled/filtered water ("unexposed") (p-value 0 is less than significant level 0.05).
v. Tap water consumption group has a higher estimated prevalence of anemia at week 9 (0.648) than bottled/filtered water consumption group (0.26).

b. [10 points] Report the point estimate of the difference in the proportion of women with anemia at week 9 of pregnancy between the tap water consumption group and the bottled/filtered water consumption group (a.k.a., the risk difference or attributable risk). Report the 95% confidence interval for the difference in proportions.

The point estimate of the difference in the proportion of women with anemia at week 9 of pregnancy between the tap water consumption group and the bottled/filtered water consumption group is 0.388. The 95% confidence interval for the mean difference is (0.31, 0.466).

c. Determine there is an association between tap or bottled/filtered water consumption and **(i)** receiving adequate prenatal care, **(ii)** smoking status prior to pregnancy, **(iii)** educational attainment, and **(iv)** parity using a chi-square test. State your statistical conclusion, p-value, and your conclusion in the context of the problem.

- **(i).** [5 points] Adequate prenatal care

```
tab2p_prenatal = table(hgb12$group_factor, hgb12$prenatal_factor,
              dnn = c("Water consumption", "Adequate prenatal care"))

tab2p_prenatal
```

```
##                          Adequate prenatal care
## Water consumption        Yes  No
##   Tap Water               95 175
##   Bottled/Filtered Water 263  52
```

```
ct_prenatal = chisq.test(tab2p_prenatal)

ct_prenatal
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab2p_prenatal
## X-squared = 140.84, df = 1, p-value < 0.00000000000000022
```

Reject H_0 and conclude H_1. That is, there is a significant association between tap or bottled/filtered water consumption and receiving adequate prenatal care (p-value = 0).

- **(ii).** [5 points] Smoking status

```
tab2p_psmoke = table(hgb12$group_factor, hgb12$psmoke_factor,
              dnn = c("Water consumption", "Smoking status"))

tab2p_psmoke
```

```
##                          Smoking status
## Water consumption        Yes  No
##   Tap Water               87 183
##   Bottled/Filtered Water  54 261
```

```
ct_psmoke = chisq.test(tab2p_psmoke)

ct_psmoke
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab2p_psmoke
## X-squared = 17.257, df = 1, p-value = 0.00003266
```

Reject H_0 and conclude H_1. That is, there is a significant association between tap or bottled/filtered water consumption and smoking status prior to pregnancy (p-value = 0).

- **(iii).** [5 points] Educational attainment

```
tab2p_ed = table(hgb12$group_factor, hgb12$ed_factor,
              dnn = c("Water consumption", "Educational attainment"))

tab2p_ed
```

```
##                          Educational attainment
## Water consumption        Less than HS HS/GHD Some college or more
##   Tap Water                       149    111                   10
##   Bottled/Filtered Water           46    238                   31
```

```
ct_ed = chisq.test(tab2p_ed)

ct_ed
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab2p_ed
## X-squared = 108.56, df = 2, p-value < 0.00000000000000022
```

Reject H_0 and conclude H_1. That is, there is a significant association between tap or bottled/filtered water consumption and educational attainment (p-value = 0).

- **(iv).** [5 points] Parity

```
tab2p_parity = table(hgb12$group_factor, hgb12$parity_factor,
              dnn = c("Water consumption", "Parity"))

tab2p_parity
```

```
##                          Parity
## Water consumption        None One Two Three or more
##   Tap Water                45 137  78            10
##   Bottled/Filtered Water   48 167  92             8
```

```
ct_parity = chisq.test(tab2p_parity)

ct_parity
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab2p_parity
## X-squared = 0.9767, df = 3, p-value = 0.8069
```

Fail to reject H_0. That is, there is not a significant association between tap or bottled/filtered water consumption and parity (p-value = 0.807).

4. [5 points] We have identified several significant differences between the two water consumption groups of interest. In particular, we found that one group tends to have a larger average decrease in hemoglobin. Since we would like to attribute this difference to herbicide exposure through drinking water, are you concerned by any of the differing

demographic/baseline characteristics between the two groups? That is, do you think differences in certain baseline characteristics will make it more difficult to attribute the differences in hemoglobin change to herbicide exposure alone?

Yes. We found that mean age, income, weight gain, week 9 hemoglobin and final hemoglobin are all significantly different in the two water consumption group from 2(c), and adequate prenatal care, smoking status and educational attainment all have significant association with water consumption. Therefore, those 8 baseline characteristics will make it more difficult to attribute the differences in hemoglobin change to herbicide exposure alone.