Week 05 Good statistical practice

Open and reproducible science: dependable computations and statistics

Homework Solutions

We will work on the data collected for the article “Avoiding overhead aversion in charity” https://www.science.org/doi/10.1126/science.1253932

The complete data that were produced are available here: https://dataverse.harvard.edu/dataverse/AvoidOHAversion

```
data_lab_experiment <-
  read.table("https://dataverse.harvard.edu/api/access/datafile/:persistentId?persistentId=doi:10.7910/
             header = TRUE, fill = TRUE)
```

```
proportions <- data_lab_experiment %>%
  mutate(overhead_level = case_when(noover == 1 ~ 0,
                                    (high == 1) | (highcover == 1) ~ 50,
                                    (low == 1) | (lowcover == 1) ~ 5),
         cover = case_when(noover == 1 ~ "control",
                           (highcover == 1) |  (lowcover == 1) ~ "covered",
                           TRUE ~ "not covered")) %>%
  group_by(overhead_level, cover) %>%
  count(allocation) %>%
  group_by(overhead_level, cover) %>%
  mutate(sum = sum(n),
         freq = n/sum(n)) %>%
  filter(allocation == 1) %>%
  dplyr::select(-allocation) %>%
  mutate(CI_lower = prop.test(n, sum)$conf.int[1],
         CI_upper = prop.test(n, sum)$conf.int[2]) %>%
  ungroup() %>%
  mutate(Treatment = c(1, 4, 2, 5, 3))
```

The article contains a lab experiment with students who were asked to decide which of two charities should receive 100 dollars. 449 participants were randomly assigned into five groups who obtained different information on charity B, information on charity A was the same in all groups.

Groups/treatments:
1. No overhead costs 2. Overhead costs of 5%, i.e. only 95 % of the donation would go to the charity
3. Overhead costs of 50%, i.e. only 50 % of the donation would go to the charity
4. Overhead costs of 5% but those would be covered by a third party
5. Overhead costs of 50% but those would be covered by a third party

In the article see page 632 third column almost at the bottom until page 633 at the top of the third column for a description and the corresponding results. See also here for the results of the experiment ordered by overhead level:

Overhead level (%)	Group	No. of donations to charity:water	Total participants	Treatment
0	control	66	90	1
5	covered	70	91	4
5	not covered	60	90	2
50	covered	65	91	5
50	not covered	43	87	3

```
proportions %>%
  dplyr::select(c(1:4,8)) %>%
  rename("Overhead level (%)" = overhead_level,
         "Group" = cover,
         "No. of donations to charity:water" = n,
         "Total participants" = sum) %>%
  kableExtra::kable() %>%
  kable_classic(full_width = F)
```

Data description

From the online supplement we know the meaning of the different variables (you only need the first seven variables):

```
# ID = random participant ID
# low = 1 = 5% overhead condition
# lowcover = 1 = 5% overhead, covered condition
# high = 1 = 50% overhead condition
# highcover = 1 = 50% overhead, covered condition
# noover = 1 = no overhead control condition
# allocation = 1 = allocated $100 to charity: water
# donbeh (On average, how often do you donate money to nonprofits?) = 1 = never to 6 = 6 or
# more times a year
# KKfamiliar (How familiar are you with Kids Korps?) = 1 = not at all to 7 = very
# CWfamiliar (How familiar are you with charity: water?) = 1 = not at all to 7 = very
# gender = 1 = male
# age = age of participant
```

Task 1 Reproduce

Reproduce and extend the results of the lab experiment.

1. Calculate the proportions

```
proportions %>%
  mutate(freq = round(freq * 100, 2),
         CI_lower = round(CI_lower * 100, 1),
         CI_upper = round(CI_upper * 100, 1)) %>%
  rename("Overhead level (%)" = overhead_level,
         "Group" = cover,
         "No. of donations to charity:water" = n,
         "Percentage (%)" = freq,
         "Total participants" = sum,
         "95% CI (lower -" = CI_lower,
         "upper)" = CI_upper,) %>%
  kableExtra::kable() %>%
  kable_classic(full_width = F) %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))
```

Overhead level (%)	Group	No. of donations to charity:water	Total participants	Percentage (%)	95% CI (lower -	upper)	Treatment
0	control	66	90	73.33	62.8	81.9	1
5	covered	70	91	76.92	66.7	84.8	4
5	not covered	60	90	66.67	55.9	76.0	2
50	covered	65	91	71.43	60.9	80.2	5
50	not covered	43	87	49.43	38.6	60.3	3

2. Report the five z-values and p-values of the article in an appropriate table

```
test1 <- prop.test(x = as.numeric(c(proportions[proportions$Treatment == 1,3],
                                    proportions[proportions$Treatment == 2,3])),
                   n = as.numeric(c(proportions[proportions$Treatment == 1,4],
                                    proportions[proportions$Treatment == 2,4])),
                   correct = FALSE)

test2 <- prop.test(x = as.numeric(c(proportions[proportions$Treatment == 1,3],
                                    proportions[proportions$Treatment == 3,3])),
                   n = as.numeric(c(proportions[proportions$Treatment == 1,4],
                                    proportions[proportions$Treatment == 3,4])),
                   correct = FALSE)

test3 <- prop.test(x = as.numeric(c(proportions[proportions$Treatment == 2,3],
                                    proportions[proportions$Treatment == 3,3])),
                   n = as.numeric(c(proportions[proportions$Treatment == 2,4],
                                    proportions[proportions$Treatment == 3,4])),
                   correct = FALSE)

test4 <- prop.test(x = as.numeric(c(proportions[proportions$Treatment == 1,3],
                                    proportions[proportions$Treatment == 5,3])),
                   n = as.numeric(c(proportions[proportions$Treatment == 1,4],
                                    proportions[proportions$Treatment == 5,4])),
                   correct = FALSE)

test5 <- prop.test(x = as.numeric(c(proportions[proportions$Treatment == 3,3],
                                    proportions[proportions$Treatment == 5,3])),
                   n = as.numeric(c(proportions[proportions$Treatment == 3,4],
                                    proportions[proportions$Treatment == 5,4])),
                   correct = FALSE)

mytests <- c("1 vs 2", "1 vs 3","2 vs 3","1 vs 5","3 vs 5")
mypvals <- c(test1$p.value,test2$p.value,test3$p.value,test4$p.value,test5$p.value)
myzvals <- c(sqrt(test1$statistic),sqrt(test2$statistic),sqrt(test3$statistic),
             sqrt(test4$statistic),sqrt(test5$statistic))


data.frame(Combination = mytests,
           "p-value" = format.pval(mypvals,1,.001),
           "z-value" = round(myzvals,2)) %>%
  kableExtra::kable() %>%
  kable_classic(full_width = F) %>%
  kable_styling(latex_options = "HOLD_position")
```

Combination	p.value	z.value
1 vs 2	0.329	0.98
1 vs 3	0.001	3.27
2 vs 3	0.020	2.32
1 vs 5	0.774	0.29
3 vs 5	0.003	3.00

3. Check the original article, do your numbers coincide? Exactly?

4. Report effect sizes and confidence intervals

```
myeffects <- c(-diff(test1$estimate),-diff(test2$estimate),-diff(test3$estimate),
               -diff(test4$estimate),-diff(test5$estimate))
mylowerCI <- c(test1$conf.int[1],test2$conf.int[1],test3$conf.int[1],
               test4$conf.int[1],test5$conf.int[1])
myupperCI <- c(test1$conf.int[2],test2$conf.int[2],test3$conf.int[2],
               test4$conf.int[2],test5$conf.int[2])


data.frame(Combination = mytests,
           Estimate = myeffects,
           "Lower CI" = mylowerCI,
           "Upper CI" = myupperCI) %>%
  kableExtra::kable(digits = 3) %>%
  kable_classic(full_width = F) %>%
  kable_styling(latex_options = "HOLD_position")
```

Combination	Estimate	Lower.CI	Upper.CI
1 vs 2	0.067	-0.067	0.200
1 vs 3	0.239	0.100	0.378
2 vs 3	0.172	0.029	0.316
1 vs 5	0.019	-0.111	0.149
3 vs 5	-0.220	-0.360	-0.080

5. Discuss the interpretation of p-values in the article. Do you find it appropriate or do you have suggestions for improvement?

Hint: for the test of proportions use `prop.test`, which also provides confidence intervals (Wilson Score Intervals, which perform quite well). The function provides `X-squared` values, the z-values of the article are the square roots of those.

Task 2 Calculate sample size

Assume that an increase in donation frequency from 65% to 75% is meaningful for a charity to consider finding coverage for overhead costs of 5%. How big would a sample need to be to detect that difference with 80% power?

Report the sample size that you calculate and compare with the sample size of the control experiment in the article.

Hint: use `power.prop.test` for a one-sided alternative. In order to obtain a sample size leave out the argument for `n` but provide all other arguments.

```
power.prop.test(p1=.75,p2=.65,power=.8,alternative="one.sided")
```

```
##
##      Two-sample comparison of proportions power calculation
##
##               n = 258.619
##              p1 = 0.75
##              p2 = 0.65
##       sig.level = 0.05
##           power = 0.8
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```