HW3\_xl2836

Xinyi Lin

10/4/2018

library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.0.0 ✔ purrr 0.2.5  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.6  
## ✔ tidyr 0.8.1 ✔ stringr 1.3.1  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ───────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(readxl)

# Problem 1

## Question 1

Assuming the mean of is .

As , we can get . As , we can get . Then,

## Question 2

# Problem 2

First, we need to import data “HeavySmoke.csv” and “NeverSmoke.csv”.

heavysmoke\_df = read\_csv(file = "./data/HeavySmoke.csv")

## Parsed with column specification:  
## cols(  
## ID = col\_integer(),  
## BMI\_base = col\_double(),  
## BMI\_6yrs = col\_double()  
## )

neversmoke\_df = read\_csv(file = "./data/NeverSmoke.csv")

## Parsed with column specification:  
## cols(  
## ID = col\_integer(),  
## BMI\_base = col\_double(),  
## BMI\_6yrs = col\_double()  
## )

## Question 1

In order to test wether BMI has changed 6 years after quitting smoking, we need to test the means of BMI\_base and BMI\_6yrs are different or not. As we don’t know the variance of two samples, we use t test.

Assuming the mean of BMI\_base is and the mean of BMI\_6yrs is , the difference between BMI\_base and BMI\_6yrs is . The samples sizes is 10.

The null hypothesis : , the alternative hypothesis : .

d\_h = heavysmoke\_df$BMI\_6yrs - heavysmoke\_df$BMI\_base # get the difference

The test statustic :

t\_s1 = mean(d\_h)\*sqrt(10)/sd(d\_h)  
t\_std1 = qt(0.975, 9)

is 4.314 and is 2.262. As larger than , we reject .

Interpreatation: At significant level, we reject and conclude that there is enough evidence to prove that the mean of BMI\_base is different from the mean of BMI\_6yrs and their BMI has changed 6 years after quitting smoking.

## Question 2

Assuming the difference between BMI\_base and BMI\_6yrs is never-smoke-group is and difference between and is . The samples sizes is 10. In order to compare the BMI changes between women that quit smoking and women who never smoked, we use two-sample independent t-test to compare the changes in two groups. First, we need to test the two samples have same variance or not.

Assuming the variances of two samples are and and the null hypothesis : , the alternative hypothesis : .

The test statistic:

d\_n = neversmoke\_df$BMI\_6yrs - neversmoke\_df$BMI\_base # get the difference  
s\_n = sd(d\_n)  
s\_h = sd(d\_h)  
f = s\_n^2/s\_h^2

As is 0.86 is smaller than 4.026 and is larger than 0.248, we fail to reject at significant level and conclude that there is no significant difference between two variances, so we used two-sample independent t-test with equal variances to compare two differences.

The null hypothesis : , the alternative hypothesis : .

The test statistic:

where is given by

s = sqrt(((10-1)\*s\_n^2 + (10-1)\*s\_h^2)/(10+10-2))  
t\_s2 = (mean(d\_n)-mean(d\_h))/(s\*sqrt(1/10+1/10))  
t\_std2 = qt(0.975, 18)

is -1.704 and is 2.101. As absolute value of smaller than , we fail to reject .

Interpreatation: At significant level, we fail to reject and conclude that there is no enough evidence to prove that the BMI changes between women that quit smoking and women who never smoked are different.

## Question 3

The corresponding 95% CI associated with the difference between changes of two groups is given by:

CI\_left = mean(d\_n)-mean(d\_h) - qt(0.975, 18)\*s\*sqrt(1/10+1/10)  
CI\_right = mean(d\_n)-mean(d\_h) + qt(0.975, 18)\*s\*sqrt(1/10+1/10)

As is 10 and is 0.05, the corresponding 95% CI associated with is [-4.041, 0.421], which means the true difference between BMI changes of women that quit smoking and women who never smoked is between -4.041 and 0.421.

## Question 4

Study design:

We can conduct a cohort study. First, we colloect the BMI of people who start to quit smoke and the BMI of them 6 years after they quited smoke. Then we select a group of 100 women from those who able to quit smoke for at least 6 years that age 50-64. After that, we select a group of 100 women that had never somke and age 50-64 in the same place as the former group and record their BMI of first and sixth years. At last, we use two sample independent t test to test whether their is difference between the changes of these two groups.

There might be some bias from selecting women in specific area so we need to find a bunch of source population and using random numbers or other ways to randomly choose 100 of them.

Sample size calculating:

We use the following formula to calculate sample sizes.

where .

And we can know that , , , .

mu\_1 = 3.0  
mu\_2 = 1.7  
sig\_1 = 2.0  
sig\_2 = 1.5  
pow\_1 = 0.8  
pow\_2 = 0.9  
signiflevel\_1 = 0.025  
signiflevel\_2 = 0.05  
  
sample\_size = function(pow, signiflevel){  
 numerator = (sig\_1^2+sig\_2^2)\*(qnorm(1-signiflevel/2)+qnorm(pow))^2  
 denominator = (mu\_1-mu\_2)^2  
 n = numerator/denominator  
 return(n)  
}

The table of sample sizes are shown below:

|  |  |  |
| --- | --- | --- |
| Choice | 2.5% significance level | 5% significance level |
| 80% power | 35 | 29 |
| 90% power | 46 | 39 |

# Problem 3

First, we need to import data “Knee.csv”.

Knee\_df = read\_csv(file = "./data/Knee.csv") %>%   
 janitor::clean\_names()

## Parsed with column specification:  
## cols(  
## Below = col\_integer(),  
## Average = col\_integer(),  
## Above = col\_integer()  
## )

## Question 1

The descriptive statistics for Below group is:

summary(Knee\_df$below)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 29 36 40 38 42 43 2

sd(Knee\_df$below, na.rm = TRUE)

## [1] 5.477226

The descriptive statistics for Average group is:

summary(Knee\_df$average)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 28.00 30.25 32.00 33.00 35.00 39.00

sd(Knee\_df$average, na.rm = TRUE)

## [1] 3.91578

The descriptive statistics for Above group is:

summary(Knee\_df$above)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 20.00 21.00 22.00 23.57 24.50 32.00 3

sd(Knee\_df$above, na.rm = TRUE)

## [1] 4.197505

According to the values calculated above, we can find that the mean and median of above group is largest and the mean and median of below group is smallest. While the variance of average group is smallest and variance of below group is largest which means the difference in condition of average group is smaller.

## Question 2

# tidy data  
Knee\_aov\_df =   
 Knee\_df %>%   
 gather(key = "group", value = "time", below:above) %>%   
 filter(!is.na(time))  
  
# get ANOVA table  
res\_knee = lm(time ~ factor(group), data = Knee\_aov\_df)   
anova(res\_knee)

## Analysis of Variance Table  
##   
## Response: time  
## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(group) 2 795.25 397.62 19.28 1.454e-05 \*\*\*  
## Residuals 22 453.71 20.62   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Assuming the mean of Below group is , the mean of Average group is and the mean of Above group is .

The null hypothesis : , the alternative hypothesis : at least two means are not equal.

According to the ANOVA table is 19.28 and is 5.72. If larger than , we reject , otherwise, we fail to reject . As larger than , we reject .

Conclusion: At significant level, we reject and conclude that at least two means of Below, Average, Above groups’ time are not equal.

## Question 3

### Bonferroni

pairwise.t.test(Knee\_aov\_df$time, Knee\_aov\_df$group, p.adj = 'bonferroni')

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: Knee\_aov\_df$time and Knee\_aov\_df$group   
##   
## above average  
## average 0.0011 -   
## below 1.1e-05 0.0898   
##   
## P value adjustment method: bonferroni

k = 2

Bonferroni adjustment:

In these case, the is 0.0033333, is 3.295. As all in t-test table are smaller than , we fail to reject and conclude there is significance different between each group.

### Tukey

aov(time ~ factor(group), data = Knee\_aov\_df) %>%   
 TukeyHSD(conf.level = 0.99)

## Tukey multiple comparisons of means  
## 99% family-wise confidence level  
##   
## Fit: aov(formula = time ~ factor(group), data = Knee\_aov\_df)  
##   
## $`factor(group)`  
## diff lwr upr p adj  
## average-above 9.428571 2.168498 16.68864 0.0010053  
## below-above 14.428571 6.803969 22.05317 0.0000102  
## below-average 5.000000 -1.988063 11.98806 0.0736833

### Dunnett

below\_g = Knee\_df$below  
average\_g = Knee\_df$average  
above\_g = Knee\_df$above  
  
DescTools::DunnettTest(list(above\_g, average\_g, below\_g), conf.level = 0.99)

##   
## Dunnett's test for comparing several treatments with a control :   
## 99% family-wise confidence level  
##   
## $`1`  
## diff lwr.ci upr.ci pval   
## 2-1 9.428571 2.520317 16.33683 0.00069 \*\*\*  
## 3-1 14.428571 7.173453 21.68369 6.9e-06 \*\*\*  
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Both of these three pairwise comparisons can test whether there is significant difference between each two groups in multiple groups. However, we can find out that the results of each test are different. The Bonferroni adjustment can test the difference between every two groups in multiple groups and the result shows that there is no significant difference in every two groups as the Bonferroni adjustment is the strictest. The Tukey adjustment can also test the difference between every two groups in multiple groups and the results shows that only below and average groups have no significant difference. The Dunnett adjustment can test the difference between reference groups(below and average groups) and the control group(above group) and the result shows that there are significant difference between both reference groups and control group.

## Question 4

At significant level, we can conclude that when seperating patients to below, average and above groups, according to Tukey and Dunnett adjustments both below and average groups are significant different from above groups in the time required in physical therapy until successful rehabilitation, which means the time required in physical therapy is associated with physical status once patients’ physical therapy is above average.

# Problem 4

## Question 1

UCBA\_df = as.tibble(datasets::UCBAdmissions) %>% # import data  
 janitor::clean\_names()  
  
admit\_male = # get the number of admitted male in each department  
 UCBA\_df %>%   
 filter(admit == "Admitted", gender == "Male")  
  
admit\_female = # get the number of admitted female in each department  
 UCBA\_df %>%   
 filter(admit == "Admitted", gender == "Female")  
  
reject\_male = # get the number of rejected male in each department  
 UCBA\_df %>%   
 filter(admit == "Rejected", gender == "Male")  
  
reject\_female = # get the number of rejected female in each department  
 UCBA\_df %>%   
 filter(admit == "Rejected", gender == "Female")  
  
x\_m = sum(admit\_male$n)  
x\_f = sum(admit\_female$n)  
  
n\_f = sum(admit\_female$n) + sum(reject\_female$n)  
n\_m = sum(admit\_male$n) + sum(reject\_male$n)  
  
p\_m = x\_m/n\_m  
p\_f = x\_f/n\_f

Using the point estimation of the proportions of female and male admitted at Berkeley. The point estimation of the proportions of female is 0.304, the point estimation of the proportions of male is 0.445.

Using the following formula to get the 95% confidence interval for the proportions of female and male:

# proportion of female  
left\_CI\_female = p\_f - qnorm(0.975)\*sqrt(p\_f\*(1-p\_f)/n\_f)  
right\_CI\_female = p\_f + qnorm(0.975)\*sqrt(p\_f\*(1-p\_f)/n\_f)  
  
# proportion of male  
left\_CI\_male = p\_m - qnorm(0.975)\*sqrt(p\_m\*(1-p\_m)/n\_m)  
right\_CI\_male = p\_m + qnorm(0.975)\*sqrt(p\_m\*(1-p\_m)/n\_m)

By using the above formula, we can get the 95% confidence interval for the proportions of female is ( 0.283, 0.325 ) and the 95% confidence interval for the proportions of male is ( 0.426, 0.464 )

According to the mean of two proportions, we can find that The point estimation of the proportions of female is lightly smaller than the point estimation of the proportions of male as well as the confidence interval which might indicates that the true proportions of female admitted in Berkeley is smaller than the true proportions of male admitted in Berkeley.

## Question 2

The null hypothesis : , the alternative hypothesis : . The test statustuc with continuity correction is given by:

when .

We create a function z.prop to calculate test statistic. The function is shown below.

z.prop = function(x1,x2,n1,n2){  
 numerator = abs((x1/n1) - (x2/n2)) - (1/(2\*n1)+1/(2\*n2))  
 p.common = (x1+x2) / (n1+n2)  
 denominator = sqrt(p.common \* (1-p.common) \* (1/n1 + 1/n2))  
 z.prop.ris = numerator / denominator  
 return(z.prop.ris)  
}  
  
z\_stat = z.prop(x\_f, x\_m, n\_f, n\_m)  
p\_stat = pnorm(z\_stat)

By calculating, we can know, is 9.571 and is 1.96. is larger than and p-value is 1.

Interpretation: At significant level, we reject and conclude that their are significant difference between the true proportions of female admitted in Berkeley and the true proportions of male admitted in Berkeley.