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 X_i is μ.

$$\begin{aligned}
E(s^2) &= E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2] \\
&= E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2\overline{x}x_i + \overline{x}^2)] \\
&= \frac{n}{n-1}E[\frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\overline{x}x_i + \overline{x}^2)] \\
&= \frac{n}{n-1}\{\frac{1}{n}E[\sum_{i=1}^{n}(x_i^2)] - 2E(\overline{x}\frac{1}{n}\sum_{i=1}^{n}x_i) + E(\overline{x}^2)\} \\
&= \frac{n}{n-1}\{\frac{1}{n}\sum_{i=1}^{n}E\,(x_i^2) - 2E(\overline{x}^2) + E(\overline{x}^2)\} \\
&= \frac{n}{n-1}\{\frac{1}{n}\sum_{i=1}^{n}E\,(x_i^2) - E(\overline{x}^2)\}
\end{aligned}$$

As $Var(x_i) = E(x_i^2) - (Ex_i)^2$, we can get $E(x_i^2) = Var(x_i) + (Ex_i)^2 = \sigma^2 + \mu^2$. As $Var(\overline{x}) = E(\overline{x}^2) - (E\overline{x})^2$, we can get $E(\overline{x}^2) = Var(\overline{x}) + (E\overline{x})^2 = \frac{\sigma^2}{n} + \mu^2$. Then,

$$\begin{aligned}
E(s^2) &= \frac{n}{n-1}[\frac{1}{n} \times n(\sigma^2 + \mu^2) - \frac{\sigma^2}{n} - \mu^2] \\
&= \frac{n}{n-1}\frac{1}{n}\sigma^2 \\
&= \sigma^2
\end{aligned}$$

Question 2

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{\overline{y}})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}^2 - 2y_{ij}\overline{\overline{y}} + \overline{\overline{y}}^2) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}^2 + \overline{\overline{y}}^2) - 2\overline{\overline{y}}\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}^2 + \overline{\overline{y}}^2) - 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\overline{\overline{y}}^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}^2 - \overline{\overline{y}}^2)
\end{aligned}$$

$$
\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_l})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y_l}-\bar{\bar{y}})^2
&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}[y_{ij}^2-2\overline{y_l}y_{ij}+(\overline{y_l})^2+(\overline{y_l})^2-2\overline{y_l\bar{\bar{y}}}+\bar{\bar{y}}^2]\\
&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}^2+\bar{\bar{y}}^2)+\sum_{i=1}^{k}\sum_{j=1}^{n_i}[2(\overline{y_l})^2-2\overline{y_l\bar{\bar{y}}}+\bar{\bar{y}}^2]\\
&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}^2+\bar{\bar{y}}^2)+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\overline{y_l}^2-2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\overline{y_l}\,y_{ij}-2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\overline{y_l}\,\bar{\bar{y}}\\
&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}^2+\bar{\bar{y}}^2)+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\overline{y_l}^2-2\sum_{i=1}^{k}(\overline{y_l}\sum_{j=1}^{n_i}y_{ij})-2\bar{\bar{y}}\sum_{i=1}^{k}(\overline{y_l}\sum_{j=1}^{n_i}1)\\
&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}^2+\bar{\bar{y}}^2)+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\overline{y_l}^2-2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\overline{y_l}^2-2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\bar{\bar{y}}^2\\
&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}^2-\bar{\bar{y}}^2)\\
&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{\bar{y}})^2
\end{aligned}
$$

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 μ_b and the mean of `BMI_6yrs` is μ_6, the difference between `BMI_base` and `BMI_6yrs` is d_h. The samples sizes n is 10.

The null hypothesis $H_0 : \mu_b = \mu_6$, the alternative hypothesis $H_1 : \mu_b \neq \mu_6$.

```
d_h = heavysmoke_df$BMI_6yrs - heavysmoke_df$BMI_base # get the difference
```

The test statustic :

$$t_{s1} = \frac{\overline{d_h} - 0}{s_d/\sqrt{n}}$$

```
t_s1 = mean(d_h)*sqrt(10)/sd(d_h)
t_std1 = qt(0.975, 9)
```

t_{s1} is 4.314 and $t_{9,0.975}$ is 2.262. As t_{s1} larger than $t_{9,0.975}$, we reject H_0.

Interpreatation: At $\alpha = 0.05$ significant level, we reject H_0 and conclude that there is enough evidence to prove that the mean of `BMI_base` μ_b is different from the mean of `BMI_6yrs` μ_6 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 d_n and difference between d_h and d_n is d. The samples sizes n 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 σ_n and σ_h and the null hypothesis $H_0 : \sigma_h = \sigma_s$, the alternative hypothesis $H_1 : \sigma_h \neq \sigma_n$.

The test statistic:

$$F = \frac{{S_1}^2}{{S_2}^2}$$

```
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 F is 0.86 is smaller than $F_{9,9,0.975}$ 4.026 and is larger than $F_{9,9,0.025}$ 0.248, we fail to reject H_0 at $\alpha = 0.05$ 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 $H_0 : d_h = d_s$, the alternative hypothesis $H_1 : d_h \neq d_n$.

The test statistic:

$$t = \frac{\overline{x_1} - \overline{x_2}}{s\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where s is given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

```
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)
```

t_{s2} is -1.704 and $t_{9,0.975}$ is 2.101. As absolute value of t_{s2} smaller than $t_{9,0.975}$, we fail to reject H_0.

Interpreatation: At $\alpha = 0.05$ significant level, we fail to reject H_0 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 d is given by:

$$\overline{x_1} - \overline{x_2} - t_{n_1+n_2-2,1-\alpha/2}s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \overline{x_1} - \overline{x_2} \leq \overline{x_1} - \overline{x_2} + t_{n_1+n_2-2,1-\alpha/2}s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

```
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 n is 10 and is 0.05, the corresponding 95% CI associated with d 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.

$$n = \frac{(\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}$$

where $\Delta = |\mu_2 - \mu_1|$.

And we can know that $\mu_1 = 3.0, \sigma_1 = 2.0, \mu_2 = 1.7, \sigma_2 = 1.5$.

```
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))
```

```r
# 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 μ_b, the mean of Average group is μ_{av} and the mean of Above group is μ_{ab}.

The null hypothesis $H_0 : \mu_b = \mu_{av} = \mu_{ab}$, the alternative hypothesis H_1 : at least two means are not equal.

According to the ANOVA table F_s is 19.28 and $F_{2,22,0.99}$ is 5.72. If F_s larger than $F_{2,22,0.99}$, we reject H_0, otherwise, we fail to reject H_0. As F_s larger than $F_{2,22,0.99}$, we reject H_0.

Conclusion: At $\alpha = 0.01$ significant level, we reject H_0 and conclude that at least two means of Below, Average, Above groups' time are not equal.

Question 3

Bonferroni

```r
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
```

```r
k = 2
```

Bonferroni adjustment:

$$\alpha^* = \frac{\alpha}{\binom{k}{2}}$$

In these case, the α^* is 0.0033333, $t_{n-k,1-\alpha^*/2}$ is 3.295. As all t in t-test table are smaller than $t_{n-k,1-\alpha^*/2}$, we fail to reject H_0 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 $\alpha = 0.05$ 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 ea
ch 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 e
ach 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 $\hat{p_f}$ is 0.304, the point estimation of the proportions of male $\hat{p_m}$ is 0.445.

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

$$\left(\hat{p} - z_{0.975}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}, \hat{p} + z_{0.975}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right)$$

```r
# 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 $\hat{p_f}$ is lightly smaller than the point estimation of the proportions of male $\hat{p_m}$ 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 $H_0 : p_f = p_m$, the alternative hypothesis $H_1 : p_f \neq p_m$. The test statustuc with continuity correction is given by:

$$z = \frac{|\hat{p_1} - \hat{p_2}| - (\frac{1}{2n_1} + \frac{1}{2n_2})}{\sqrt{\hat{p}\,\hat{q}\,(\frac{1}{n_1} + \frac{1}{n_2})}}$$

when $\hat{p} = \frac{n_1\hat{p_1} + n_2\hat{p_2}}{n_1 + n_2}$.

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

```r
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,z is 9.571 and $z_{1-\alpha/2}$ is 1.96. z is larger than $z_{1-\alpha/2}$ and p-value is 1.

Interpretation: At $\alpha = 0.05$ significant level, we reject H_0 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.