HW3

Zhuodiao Kuang

2023-10-25

```
library(MASS)
library(dplyr)
```

Problem 1

Some medical professionals claim that the average weight of American women is 171 pounds. The column lwt holds the mother's weight (in pounds) at last menstrual period, i.e. her pre-pregnancy weight. Use this column for the following questions.

a) Construct a 95% confidence interval of true mean weight of American women

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)$$

So, the interval is

$$\bar{X} - \frac{s}{\sqrt{n}} t_{0.975} \leq \mu \leq \bar{X} + \frac{s}{\sqrt{n}} t_{0.975}$$

```
wt <- birthwt |> pull(lwt)
n <- length(wt)
df = n-1
t <- qt(0.975,df = df)
lq <- mean(wt) - sd(wt)/sqrt(n)*t
uq <- mean(wt) + sd(wt)/sqrt(n)*t
print(c(lq,uq))
```

```
[1] 125.4270 134.2027
```

So, the 95% confidence interval of true mean weight of American women is $[125.4270, 134.2027]$.

b) Interpret the confidence interval. We are 95% confident that the average weight of all American women in the population is between 125.4270 and 134.2027 pounds. This means that if we repeated the same sampling procedure many times and calculated a confidence interval for each sample, about 95% of these intervals would contain the true population mean weight of American women.

c) Comment on the validity of the statement above ("Some medical professionals claim that the average weight of American women is 171 pounds"). In other words, what can we say about this statement given our confidence interval from part a? Given our confidence interval of [125.4270 ,134.2027] for the true mean weight of American women, we can say that the statement that the average weight of American women is 171 pounds is very unlikely to be true. This is because 171 pounds is far outside the range of plausible values for the population mean weight of American women based on our sample data. If the statement were true, it would mean that our sample was not representative of the population or that there was a large sampling error or bias in our data collection process. Therefore, we have strong evidence to reject the claim that the average weight of American women is 171 pounds.

Problem 2

In this data set, we have a variable (smoke) indicating the smoking status of the mothers during pregnancy. Some doctors believe that smoking status is related to weight. Using the columns smoke and lwt, test this claim. (Note: a value of 1 indicates the mother is in the "smoking" group.)

a) Test for the equality of variances between the two groups. (Use a 5% significance level.) Testing the hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2 vs H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1} under H_0$$

```
wt1 <- birthwt |> filter(smoke == 1) |> pull(lwt)
wt0 <- birthwt |> filter(smoke == 0) |> pull(lwt)
n1 <- length(wt1)
n0 <- length(wt0)
s1 <- var(wt1)
s0 <- var(wt0)
F = s1/s0

lq <-qf(0.025,n1-1,n0-1)
uq <-qf(0.975,n1-1,n0-1)

F;lq;uq
```

```
[1] 1.412636
```

```
[1] 0.6518345
```

```
[1] 1.50466
```

Because F is equal to 1.41, it's between the values, $F_{n_1-1,n_2-1,\alpha/2}$ and $F_{n_1-1,n_2-1,1-\alpha/2}$. So, we cannot reject H_0 and conclude that there is no evidence to support that the two population variances are not equal.

b) Given your answer from part a, what kind of hypothesis test will you perform? Given that two population variances are equal, we can test the hypothesis that both groups have the same mean(two-sided).

Testing the hypothesis:

$$H_0 : \mu_1 = \mu_2 vs H_1 : \mu_1 \neq \mu_2$$

c) Conduct your chosen hypothesis test from part b at the 10% significance level. What is your decision regarding the null? Interpret this result in the context of the problem.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}, \sim t_{n_1+n_2-2} under H_0$$

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

```
s_2 <- (var(wt1)*(n1-1)+var(wt0)*(n0-1))/(n1+n0-2)
s <-sqrt(s_2)
t <- (mean(wt1) - mean(wt0) )/(s*sqrt(1/n1+1/n0))
t0<-qt(0.95,n1+n0-2)
abs(t);t0
```

```
[1] 0.6047303
```

```
[1] 1.653043
```

$$\because |t| = 0.60 \leq t_{n_1+n_2-2,1-\alpha/2} = 1.65, \alpha = 0.10$$

So, we cannot reject the null hypothesis that the hypothesis that both groups have the same mean.

Problem 3

According to the CDC, approximately 20% of pregnant American women suffer from hyper-tension. Do our data support this claim? (Use column ht - a value of 1 means the mother is suffering from hypertension.)

a) Conduct a 99% confidence interval and interpret the results. What can we conclude about the CDC's claim from this interval? So, the interval is

$$(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} - \frac{1}{2n}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + \frac{1}{2n})$$

```
ht <- birthwt |>pull(ht)
p <- mean(ht)
n <- length(ht)
s <- sqrt(p*(1-p)/n)
z <- qnorm(0.995)
lq <- p-s*z-1/(2*n)
uq <- p+s*z+1/(2*n)
lq;uq
```

```
[1] 0.01515862
```

```
[1] 0.1118255
```

So, the 99% confidence interval is [0.015,0.112], and 0.2 is not contained in the interval. That means we are 99% confident that the percent of pregnant American women suffer from hyper-tension is between 1.5% and 11.2%. This means that if we repeated the same sampling procedure many times and calculated a confidence interval for each sample, about 99% of these intervals would contain the true population mean percent of pregnant American women suffer from hyper-tension.

Since 20% is outside the 99% confidence interval, it is very unlikely that the 20% is the true population mean.

b) Conduct a one-sided hypothesis test at the = 0.1 level. In this test, we want to see if the true proportion is indeed less than the claimed 20%. What can we conclude about the CDC's claim?

$$H_0 : \mu \geq 0.2 H_1 : \mu < 0.2$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1), under H_0$$

$$p_0 = 0.2, \alpha = 0.9$$

```
p0 <- 0.2
z <- (p-p0)/sqrt(p0*(1-p0)/n)
zalpha <- qnorm(0.9)
abs(z);zalpha
```

```
[1] 4.691685
```

```
[1] 1.281552
```

Even we assume $p_0 = 0.2$ for the test, $|z| > z_{1-\alpha/2}$, so we reject H_0 and think that the true proportion is indeed less than the claimed 20%. So, based on the result of the test, we conclude that it's very unlikely that the proportion exceeds 20%, CDC's claim may be too exaggerating.

Problem 4

Is there a difference between uterine irritability in the group of pregnant women who smoke vs the group of pregnant women that don't smoke? (Use columns ui and smoke.) Conduct a hypothesis test at the = 0.01 level. What can we conclude about the proportions of women with uterine irritability between the smoking groups? Suppose we test:

$$H_0 : p_1 = p_2 = p H_1 : p_1 \neq p_2$$

If the null hypothesis is true:

$$\hat{p_1} - \hat{p_2} \sim N(0, p(1-p)(\frac{1}{n_1} + \frac{1}{n_2}))$$

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

So, the test statistic is given by:

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

```
ui1 <- birthwt |> filter(smoke == 1) |> pull(ui)
ui0 <- birthwt |> filter(smoke == 0) |> pull(ui)
p1 <-mean(ui1)
p0 <-mean(ui0)
n1 = length(ui1)
n0 = length(ui0)
p<-(n1*p1+n0*p0)/(n1+n0)
z = (p1-p0)/sqrt(p*(1-p)*(1/n1+1/n0))
abs(z);qnorm(0.995)
```

```
[1] 0.8545449
```

```
[1] 2.575829
```

At 0.01 significance level, |z-stat| < 2.58; thus we don't reject the null and conclude that there is no significant difference between the proportions of women with uterine irritability between the smoking groups.

Problem 5

Is race related to birth weight? (Use columns race and bwt.)

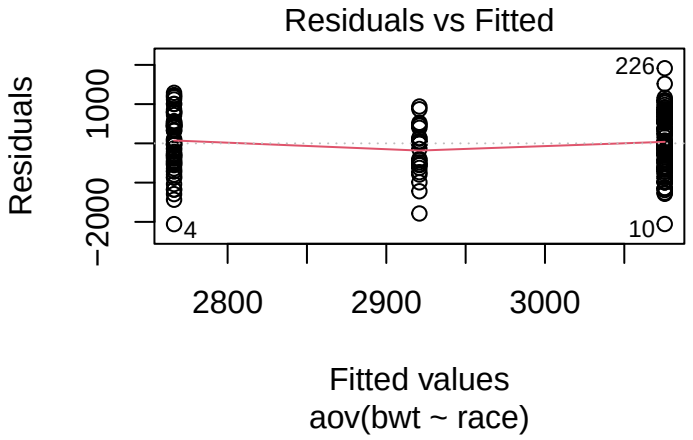
a) What test would be most appropriate to answer this question? ANOVA. Since we want to compare the differences between 3 birth weight groups.

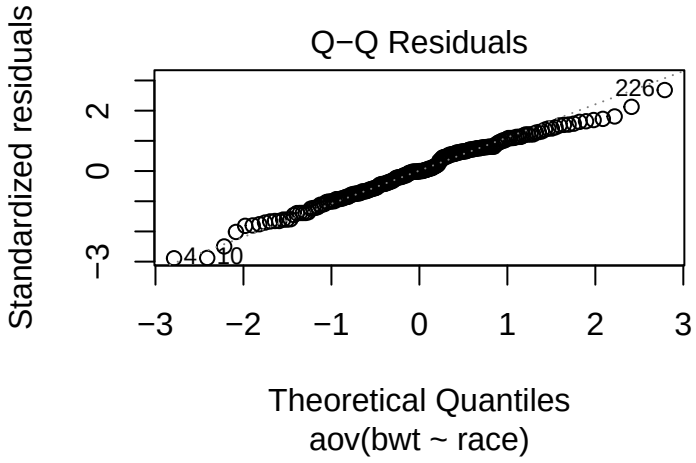
b) What assumptions are we making by using this test? Are all assumptions met? Assumptions: 1. There are k populations of interest(k>2); 2. Samples are drawn independently from the underlying populations; 3. Homoscedasticity - the variances of k populations area equal; 4. Normality - the distribution of the error terms is normal.

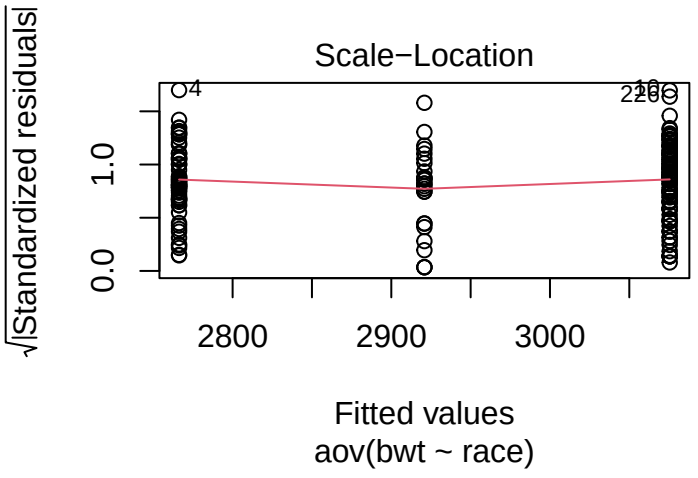
```
data = birthwt
# Perform the ANOVA test
model <- aov(bwt ~ race, data = data)

# Find the best-fit model
summary(model)
```

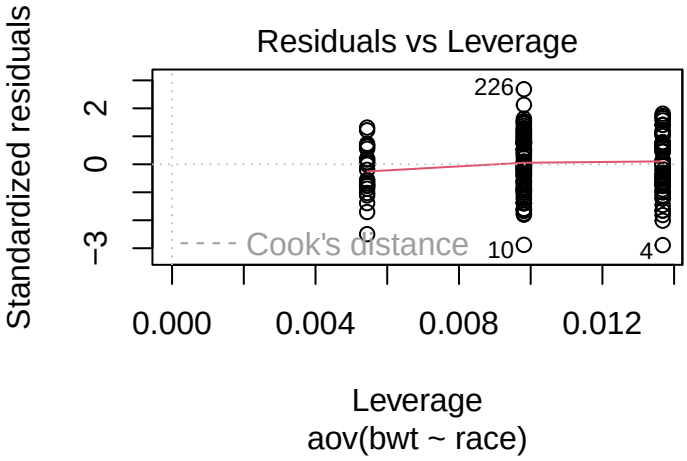
```
             Df   Sum Sq Mean Sq F value  Pr(>F)
race          1  3790184 3790184   7.369 0.00726 **
Residuals   187 96179472  514329
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Check for homoscedasticity
plot(model)
```


Residuals vs Fitted
Residuals
−2000
1000
2800
2900
3000
226
4
10
Fitted values
aov(bwt ~ race)


Q−Q Residuals
Standardized residuals
−3
0
2
−3
−2
−1
0
1
2
3
226
4
10
Theoretical Quantiles
aov(bwt ~ race)


Scale−Location
√|Standardized residuals|
0.0
1.0
2800
2900
3000
4
226
Fitted values
aov(bwt ~ race)



Conclusion: 1. There are 3 populations of interest; 2. We are not sure about whether samples are drawn independently from the underlying populations; From the plot I draw, I think homoscedasticity and normality are satisfied.

c) Conduct the test at the 5% significance level and interpret your results. Be sure to write the hypothesis you are testing. $H_0 : \mu_1 = \mu_2 = \mu_3 vs. H_1$:at least two means are not equal

$$F = \frac{BetweenSS/(k-1)}{WithinSS/(n-k)} \sim F_{k-1,n-k}$$

From Problem b):

$$\because F_{k-1,n-k,1-\alpha} = 3.89, \alpha = 0.05$$

$$\because F = 7.369 > 3.89$$

So, we reject the null hypothesis and think that there are at least two weight groups of different races are different.

d) Perform multiple comparisons - which races are significantly different? Interpret your results. Given that two population variances are equal, we can test the hypothesis that both groups have the same mean(two-sided).($\alpha = 0.05$)

Testing the hypothesis:

$$H_0 : \mu_1 = \mu_2 vs H_1 : \mu_1 \neq \mu_2$$
$$H_0 : \mu_1 = \mu_3 vs H_1 : \mu_1 \neq \mu_3$$

$$H_0 : \mu_2 = \mu_3 vs H_1 : \mu_2 \neq \mu_3$$

Just take Group1 and Group2 as an example:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}, \sim t_{n_1+n_2-2} under H_0$$

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

```
wt1 <- birthwt |> filter(race == 1) |> pull(bwt)
wt2 <- birthwt |> filter(race == 2) |> pull(bwt)
wt3 <- birthwt |> filter(race == 3) |> pull(bwt)
n1 <- length(wt1)
n2 <- length(wt2)
n3 <- length(wt3)
s_2 <- (var(wt1)*(n1-1)+var(wt2)*(n2-1))/(n1+n2-2)
s <-sqrt(s_2)
t <- (mean(wt1) - mean(wt2) )/(s*sqrt(1/n1+1/n2))
t0<-qt(0.975,n1+n2-2)
abs(t);t0
```

```
[1] 2.43935
```

```
[1] 1.97993
```

$$\because |t| = 2.44 \geq t_{n_1+n_2-2,1-\alpha/2} = 1.98, \alpha = 0.05$$

So, we reject the null hypothesis that the hypothesis that both groups have the same mean of the baby birth weight. (Race1 and Race2)

```
s_2 <- (var(wt1)*(n1-1)+var(wt3)*(n3-1))/(n1+n3-2)
s <-sqrt(s_2)
t <- (mean(wt1) - mean(wt3) )/(s*sqrt(1/n1+1/n3))
t0<-qt(0.975,n1+n3-2)
abs(t);t0
```

```
[1] 2.57513
```

```
[1] 1.974808
```

$$\because |t| = 2.58 \geq t_{n_1+n_2-2,1-\alpha/2} = 1.97, \alpha = 0.05$$

So, we reject the null hypothesis that the hypothesis that both groups have the same mean of the baby birth weight. (Race1 and Race3)

```
s_2 <- (var(wt2)*(n2-1)+var(wt3)*(n3-1))/(n2+n3-2)
s <-sqrt(s_2)
t <- (mean(wt2) - mean(wt3) )/(s*sqrt(1/n2+1/n3))
t0<-qt(0.975,n2+n3-2)
abs(t);t0
```

```
[1] 0.5290079
```

```
[1] 1.986377
```

$$\because |t| = 0.53 \leq t_{n_1+n_2-2,1-\alpha/2} = 1.99, \alpha = 0.05$$

So, we cannot reject the null hypothesis that both groups have the same mean of the baby birth weight. (Race1 and Race3)

Conclusion The mean of birth weight of Race1 is significantly different from other two groups.