p8130_hw3_zj2357

Zekai Jin(zj2357)

2022-10-16

Including the data and packages used for the homework

```
library(tidyverse)
library(MASS)
knitr::opts_chunk$set(
  fig.width = 6,
  fig.asp = .6,
  out.width = "90%"
)
birthwt = birthwt
```

Problem 1

a)
Because the sample size n=189 is large enough and the sample mean is unknown, the sample mean follows t distribution. Thus, the 95% confidence interval is

$$(\bar{X} - t_{188,0.975}\frac{s}{\sqrt{n}}, \bar{X} + t_{188,0.975}\frac{s}{\sqrt{n}})$$

This can be calculated using the following R code

```
n = nrow(birthwt)
x_b=mean(birthwt$lwt)
x_s=sd(birthwt$lwt)
t_a=qt(.975,df = 188)
CI=c(x_b-t_a*x_s/sqrt(n),x_b+t_a*x_s/sqrt(n))
CI
```

```
## [1] 125.4270 134.2027
```

Thus, the 95% CI is $(125.4, 134.2)$

b)
Using the similar calculation, 95% of all CIs will include the true mean weight of American women at last menstrual period.

c)
Because the confidence interval doesn't contain 171 pounds, based on our data, we can say that the true mean of our sampled weight of American women is not 171 pounds.

However, since the data is only collected in Baystate Medical Center during 1986, which is not representative of all American women at present. Thus, we cannot refuse their claim.

Problem 2

a)
Denote group 1 to be smoking group and group 0 to be the other group.

We suggest that

$$H_0 : \sigma_1 = \sigma_0$$
$$H_1 : \sigma_1 \neq \sigma_0$$

Under H_0 , the statistic

$$F = \frac{s_1^2}{s_0^2} \sim F_{n_1-1,n_0-1}$$

Calculate the mean, sd and number of each group using the following code:

```
birthwt %>%
  group_by(smoke) %>%
  summarize(
    n=n(),
    sd=sd(lwt),
    mean = mean(lwt)
  )
```

```
## # A tibble: 2 x 4
##   smoke     n    sd  mean
##   <int> <int> <dbl> <dbl>
## 1     0   115  28.4  131.
## 2     1    74  33.8  128.
```

Because

$$F = \frac{33.786^2}{28.427^2} = 1.412 \in (F_{73,114,0.025}, F_{73,114,0.975}) = (0.652, 1.505)$$

Thus, we cannot reject that the Variance between smoking and non smoking group is same.

b)
I will use two-sample t-test under equal variance assumption.

c)
Calculate the t statistics using the following code:

```
var_est = (28.427^2*(115-1)+33.787^2*(74-1))/(115+74-2)
t = (128.135-130.896)/sqrt(var_est*(1/115+1/74))
t
```

```
## [1] -0.6048338
```

Because

$$|t| = 0.60 < t_{187,0.95} = 1.65$$

we cannot say that smoking is related to weight in our sampled population.

Problem 3

a)
Calculate the number of hyper-tension women in our sample:

```
birthwt %>%
  group_by(ht) %>%
  summarise(n=n())
```

```
## # A tibble: 2 x 2
##      ht     n
##   <int> <int>
## 1     0   177
## 2     1    12
```

Based on our result,

$$\hat{p} = \frac{12}{189} = 0.0635$$

Because $\hat{p}$ is small, we chose not to use normal approximation.

To calculate the CI, use the method provided by R:

```
p_est=12/189
n=189
CI=c(qbinom(0.005,n,p_est)/n,qbinom(0.995,n,p_est)/n)
CI
```

```
## [1] 0.02116402 0.11111111
```

Thus, the 99% CI for our dataset is $(0.021, 0.111)$.

Because 0.2 is not in our interval, our data refuses CDC's claim.

However, similar to Problem 1, our dataset is not randomly sampled in all US pregnant woman. Thus, the result should be intepreted with caution.

b)
Suggest that

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

Because

$$n\hat{p}(1-\hat{p}) = 30.24 > 5$$

we can use normal approximation.

Under H_0, the test statistic

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = -4.96$$

Because

$$|z| = 4.96 > z_{0.9} = 1.28$$

Refuse H_0. The hypertension proportion in our sampled population is less than 20% under $a = 0.1$

Problem 4

First, we should summarize the data grouped on smoking

```
birthwt %>%
  group_by(smoke) %>%
  summarise(
    n=n(),
    n_ui = sum(ui),
    p = n_ui/n
  )
```

```
## # A tibble: 2 x 4
##   smoke     n  n_ui     p
##   <int> <int> <int> <dbl>
## 1     0   115    15 0.130
## 2     1    74    13 0.176
```

because

$$n_0\hat{p_0}(1-\hat{p_0}) = 13 > 5$$
$$n_1\hat{p_1}(1-\hat{p_1}) = 10.73 > 5$$

use z-test for proportion:

$$H_0 : p_1 = p_0$$
$$H_1 : p_1 \neq p_0$$

Thus,

$$\hat{p} = \frac{n_{ui}}{n} = 0.148$$
$$z = \frac{\hat{p_1} - \hat{p_0}}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_0)}} = 0.855$$

Because $|z| = 0.855 < z_{0.995} = 2.576$

We cannot reject there are difference in uterine irritability between smoking groups under $a = 0.01$.

Problem 5

a)
ANOVA. Because we want to compare levels between more than 2 groups.

b)
1. There are 3 populations of interest
2. The samples are drawn independently from the populations
3. The variances of the populations are equal
4. The error terms are normal

The term 1 and 2 are met. The term 4 is met based on central limit theorom. To test term 3, group the data based on race:

```
birthwt %>%
  group_by(race) %>%
  summarise(
    mean = mean(bwt),
    sd = sd(bwt),
    n= n(),
    sum=sum(bwt)
  )
```

```
## # A tibble: 3 x 5
##    race  mean    sd     n    sum
##   <int> <dbl> <dbl> <int>  <int>
## 1     1 3103.  728.    96 297861
## 2     2 2720.  639.    26  70712
## 3     3 2805.  722.    67 187954
```

Based on the table above, conduct pariwise F test for equality of variance:

$$F_{1,2} = \frac{728^2}{639^2} = 1.30 < F_{94,25,0.95} = 1.78$$

$$F_{3,2} = \frac{722^2}{639^2} = 1.28 < F_{94,25,0.95} = 1.81$$

$$F_{1,3} = \frac{728^2}{722^2} = 1.02 < F_{94,66,0.95} = 1.46$$

Thus, we can suggest equal variance for our population. Term 3 is met.

c)
Suggest that

$$H_0 : u_1 = u_2 = u_3$$
$$H_1 : Otherwise$$

Under H_0, we have

$$\text{Between SS} = \sum_{i=1,2,3} n_i \bar{y_i}^2 - \frac{y_{..}^2}{n} = 4383107$$

$$\text{Within SS} = \sum_{i=1,2,3} (n_i - 1) s_i^2 = 94961249$$

$$F = \frac{\text{Between SS}/2}{\text{Within SS}/186} = 4.29 > F_{2,186,0.95} = 3.04$$

Thus, reject H_0. Race is related to birth weight.

d)
According to Bonferroni adjustment,

$$a^* = \frac{a}{\binom{3}{2}} = 0.017$$

Because we suggest equal variance,

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_3^2 = (n_3 - 1)s_3^2}{n - 3}} = 714.5$$

Thus, the t-statistics for the 3 groups is calculated as follows

$$t_{1,2} = \frac{\bar{X_1} - \bar{X_2}}{s\sqrt{1/n_1 + 1/n_2}} = 2.42$$

$$t_{2,3} = \frac{\bar{X_1} - \bar{X_2}}{s\sqrt{1/n_1 + 1/n_2}} = -0.51$$

$$t_{1,3} = \frac{\bar{X_1} - \bar{X_2}}{s\sqrt{1/n_1 + 1/n_2}} = 2.62$$

Because

$$t_{120,0.983} = 2.14 < |t_{1,2}|$$

$$t_{161,0.983} = 2.14 < |t_{1,3}|$$

There is significant difference in population mean between race 1 and race 2,3 under $a = 0.05$.

No significant difference between race 2 and 3 is found.