p8130_hw3_yl5508

Yifei LIU (yl5508)

2023/10/27

```
library(tidyverse)
library(MASS)
data("birthwt")
birth_data = as_tibble(birthwt)
```

Problem 1

(a)

```
#one-sample t-test, unknown variance
t_lower = qt(0.025,188)
t_upper = qt(0.975,188)
print(paste("(", mean(birth_data$lwt)+t_lower*sd(birth_data$lwt)/sqrt(nrow(birth_data)), ",",
            mean(birth_data$lwt)+t_upper*sd(birth_data$lwt)/sqrt(nrow(birth_data)), ")"))
```

```
## [1] "( 125.426976563035 , 134.202653066594 )"
```

Treat data `lwt` from `birthwt` dataset as a sample:
- Sample size: 189
- Sample mean: 129.8148148
- Sample sd: 30.5793804

For true from population is unknown, we need to calculate the estimated standard error: $\frac{s}{\sqrt{n}} = \frac{30.6}{\sqrt{189}} = 2.2$.
For two-sided test, $t_{n-1,1-\alpha/2} = 1.97$.
So, the confidence interval(CI) for the population mean weight of American women is $125.4 < \mu < 134.2$.

(b) We are 95% confident that the true population mean lies between the lower (125.4) and the upper (134.2) limits of the interval.
OR: Over the collection of all 95% confidence intervals that could be constructed from repeated samples of size n (189), 95% of them will contain the true population mean.

(c) Because the population mean(171) doesn't lie in the confidential interval(125.4< <134.2), we can say that the medical claim about average weight of American women is not true (probability of which is less than 5%), or the sample we have cannot correctly reflect the character of population.

Problem 2

(a)

```r
lwt_smo =
  birth_data |>
  filter(smoke == 1) |>
  dplyr::select(lwt)

lwt_nsm =
  birth_data |>
  filter(smoke == 0) |>
  dplyr::select(lwt)
```

- Sample size: 74(smoking group) and 115(non-smoking group)
- Sample mean: 128.1351351(smoking group) and 130.8956522(non-smoking group)
- Sample sd: 33.7867301(smoking group) and 28.4269991(non-smoking group)

Test for equality of variances.
Testing the hypotheses: $H_0 : \sigma_1^2 = \sigma_2^2,\ H_1 : \sigma_1^2 \neq \sigma_2^2$
With = 0.05, compute the test statistic: $F = \frac{s_1^2}{s_2^2} = \frac{33.8^2}{28.4^2} = 1.4$
Critical value: $F_{73,114,0.975}$=1.5046602, $F_{73,114,0.025}$=0.6518345
Reject H_0: if $F_{stat} < F_{n_1-1,\ n_2-1,\ 1-\alpha/2}$ or $F_{stat} > F_{n_1-1,\ n_2-1,\ \alpha/2}$
Fail to reject H_0: if $F_{n_1-1,\ n_2-1,\ \alpha/2} \leq F_{stat} \leq F_{n_1-1,\ n_2-1,\ 1-\alpha/2}$
Cause $0.7 < F_{stat} < 1.5$, we fail to reject the null hypothesis, meaning we do not have evidence that the variances are unequal between smoking group and non-smoking group.

```r
#or we can use R code.
var.test(lwt ~ smoke, data = birth_data, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  lwt by smoke
## F = 0.7079, num df = 114, denom df = 73, p-value = 0.09744
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4614313 1.0651436
## sample estimates:
## ratio of variances
##          0.7078964
```

```r
#p-value>0.05, fail to reject H0.
```

(b) Given that we fail to consider the variances as equal, we shall use 2 independt samples t-test for unknown population variance with equal sample variances.

(c)
Testing the hypotheses: $H_0 : \mu_1 = \mu_2,\ H_1 : \mu_1 \neq \mu_2$
Compute: $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{73 \cdot 33.8^2+114 \cdot 28.4^2}{74+115-2} = 938.3$ With = 0.1, compute the test statistic:

$t = \frac{\overline{X_1}-\overline{X_2}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{128.1-130.9}{\sqrt{938.3}\cdot\sqrt{\frac{1}{74}+\frac{1}{115}}} = -0.6$
Critical value: $t_{n_1+n_2-2,1-\alpha/2} = t_{187,0.95}$=1.6530429
Reject H_0: if $|t| > t_{n_1+n_2-2,1-\alpha/2}$
Fail to reject H_0: if $|t| \leq t_{n_1+n_2-2,1-\alpha/2}$
Cause $|t_{stat}| = 0.6 < 1.7$, we fail to reject the null hypothesis, meaning we do not have evidence that the mean number is different between smoking group and non-smoking group.

```
#or we can use R code.
t.test(lwt ~ smoke, data = birth_data, alternative = "two.sided", conf.level = 0.9, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  lwt by smoke
## t = 0.60473, df = 187, p-value = 0.5461
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 90 percent confidence interval:
##  -4.785414 10.306448
## sample estimates:
## mean in group 0 mean in group 1
##        130.8957        128.1351
```

```
#p-value>0.05, fail to reject H0.
```

Problem 3

(a)

```
lwt_hyp =
  birth_data |>
  filter(ht == 1) |>
  dplyr::select(lwt)

lwt_nhy =
  birth_data |>
  filter(ht == 0) |>
  dplyr::select(lwt)
```

- Sample size: 12(hypertension group) and 177(non-hypertension group)
- Sample mean: 157.5(hypertension group) and 127.9378531(non-hypertension group)
- Sample sd: 47.0348034(hypertension group) and 28.3687484(non-hypertension group)

From data shown above, $\hat{p} = \frac{12}{189} = 0.06$
A 99% confidence interval for one population proportion is given by: $(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$,
i.e. $(0.06 - z_{0.995}\sqrt{\frac{0.06(1-0.06)}{189}}, 0.06 + z_{0.995}\sqrt{\frac{0.06(1-0.06)}{189}})$=(0.018,0.109)
Interpretation: We are 99% confident that the true population proportion lies between the lower (0.018)

and the upper (0.109) limits of the interval. The given 20% proportion is out of such interval, so we shall reject the hypothesis at the =0.1 level that CDC's claim is not true.

```
#or use R code.
prop.test(nrow(lwt_hyp),nrow(birth_data),p = 0.2, alternative = "two.sided", conf.level = 0.99)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  nrow(lwt_hyp) out of nrow(birth_data), null probability 0.2
## X-squared = 21.167, df = 1, p-value = 4.21e-06
## alternative hypothesis: true p is not equal to 0.2
## 99 percent confidence interval:
##  0.02926609 0.12892679
## sample estimates:
##          p
## 0.06349206
```

(b)
Testing the hypotheses: $H_0 : p = p_0,\ H_1 : p < p_0$
With $= 0.1$, compute the test statistic: $z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.06-0.20}{\sqrt{0.2(1-0.2)/189}} = -4.8$
Critical value: $z_\alpha = z_{0.1}$=-1.2815516
Reject H_0: if $z < z_\alpha$
Fail to reject H_0: if $z \geq z_\alpha$
Cause $z_{stat} = -4.8 < -1.3$, we would reject the null hypothesis at the =0.1 level, meaning we have evidence that the true proportion is less than what CDC claims is.

```
#or use R code.
prop.test(nrow(lwt_hyp),nrow(birth_data),p = 0.2, alternative = "less", conf.level = 0.9)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  nrow(lwt_hyp) out of nrow(birth_data), null probability 0.2
## X-squared = 21.167, df = 1, p-value = 2.105e-06
## alternative hypothesis: true p is less than 0.2
## 90 percent confidence interval:
##  0.00000000 0.09324317
## sample estimates:
##          p
## 0.06349206
```

```
#p-value<0.1, reject H0.
```

Problem 4

```
ui_smo =
  birth_data |>
  filter(smoke == 1) |>
  dplyr::select(ui)

ui_nsm =
  birth_data |>
  filter(smoke == 0) |>
  dplyr::select(ui)
```

- Sample size: 74(smoking group) and 115(non-smoking group)
- Sample proportion: 0.3915344(smoking group) and 0.6084656(non-smoking group)

Testing the hypotheses: $H_0 : p_1 = p_2,\ H_1 : p_1 \neq p_2$
$\hat{p} = \frac{n_1\hat{p_1}+n_2\hat{p_2}}{n1+n2} = \frac{74 \cdot 0.39+115 \cdot 0.61}{189} = 0.52$ With $= 0.01$, compute the test statistic: $z = \frac{\hat{p_1}-\hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{0.39-0.61}{\sqrt{0.52(1-0.52)(\frac{1}{74}+\frac{1}{115})}} = -2.95$
Critical value: $z_{1-\alpha/2} = z_{0.995}$=2.5758293
Reject H_0: if $|z| > z_{1-\alpha/2}$
Fail to reject H_0: if $|z| \leq z_{1-\alpha/2}$
Cause $|z_{stat}| = 2.95 > 2.58$, we would reject the null hypothesis at the $=0.1$ level, meaning we have evidence that the proportion of women with uterine irritability is different between smoking group and non-smoking group.

```
#or we can use R code.
prop.test(c(birth_data|>group_by(ui,smoke)|>summarise(count=n())|>filter(smoke==1, ui==1)|>pull(), birth
```

```
## `summarise()` has grouped output by 'ui'. You can override using the `.groups`
## argument.
## `summarise()` has grouped output by 'ui'. You can override using the `.groups`
## argument.

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(pull(filter(summarise(group_by(birth_data, ui, smoke), count = n()), smoke == 1, ui == 1)),
## X-squared = 0.41576, df = 1, p-value = 0.5191
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  -0.05509934  0.14558112
## sample estimates:
##    prop 1    prop 2
## 0.1756757 0.1304348
```

```
#p-value>0.05, fail to reject H0.
```

Problem 5

(a)
ANOVA: test for any differences in mean response among different levels of a factor.

(b)

```
bartlett.test(birth_data$bwt, birth_data$race)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  birth_data$bwt and birth_data$race
## Bartlett's K-squared = 0.65952, df = 2, p-value = 0.7191
```

Assumption:
- There are k population of interest (k 2). There are 3 races here.
- The samples are drawn independently from the underlying populations. Samples are picked independently.
- Homoscedasticity: the variance of the k populations are equal, which means variance of the outcome does not depend on the sample. Samples in different groups share the same variance (p-value>0.05).
- Normality: the distribution of the error terms are normal. Cause n=189, the sampling distribution would be normal.

(c)

Testing the hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3,\ H_1 : \textit{at least two means are not equal}$

```
aov(bwt ~ race, data = birth_data) |>
  summary()
```

```
##              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
```

Cause p-value=0.007<0.05, we would reject the null hypothesis at the =0.05 level, meaning we have evidence that at least two races have different mean birth weight.

(d)

```
pairwise.t.test(birth_data$bwt, birth_data$race,
                p.adjust.method = "bonf", paired = FALSE, alternative = "two.sided")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
```

```
## data:  birth_data$bwt and birth_data$race
##
##   1     2
## 2 0.049 -
## 3 0.029 1.000
##
## P value adjustment method: bonferroni
```

The adjusted p-value between group 1 and 2, group 1 and 3 are less than 0.05, which means the mean birth weight of group 1 is different from group 2 and 3 with significance level =0.05.
However, the adjusted p-value between group 2 and 3 are greater than 0.05, which means under significance level =0.05 there is no evidence showing that there is a difference in mean birth weight between the two groups.