HW3

Pei Tian, pt2632

2023-10-08

```
library(MASS)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
data("birthwt")
birthwt <- tibble(birthwt)
```

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.

$\hat{\mu} = \overline{lwt} = 129.8148$

$s = std(lwt) = 30.57938$

Denote that the distribution of column `lwt` subject to X, then $\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1} = t_{188}$.

So the 95% confidence interval of $\overline{X}$ is $(\hat{\mu} - t_{188,0.975}\frac{s}{\sqrt{n}}, \hat{\mu} + t_{188,0.975}\frac{s}{\sqrt{n}})$, which equals to $(125.4271, 134.2025)$.

```
# mean
u = mean(pull(birthwt, lwt))
# degree of freedom
df = nrow(birthwt)
# standard error
se = sd(pull(birthwt, lwt)) / sqrt(nrow(birthwt))
# 95% two-sided CI
c(u - qt(0.975, df) * se, u + qt(0.975, df) * se)
```

```
## [1] 125.4271 134.2025
```

b) Interpret the confidence interval.

Interpretation: We are 95% confident that the true mean weight of American women lies between 125.4552 and 134.1744.

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?

In the prespective of `birthwt` dataset, the statement that the average weight of American women is 171 pounds is invaild in 95% confidence level because 171 doesn't lies in 95% confidence interval. However, given that this dataset only includes weight data of women in last menstrual period, this conclusion about validity is not convinced enough because the sample is selected for specific scenario, which is not representative enough for whole American women population.

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.)

Denote variance of smoking mother population is σ_1^2 and variance of non-smoking mother population is σ_2^2, variance of smoking mother sample is s_1^2 and variance of non-smoking mother sample is s_2^2.

Test hypothesis : $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$,then $F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1} = F_{73,114}$.

At 5% significance level, when $F < F_{73,114,0.025}$ or $F > F_{73,114,0.975}$, which means $F < 0.6518345$ or $F > 1.5046602$, we reject the null hypothesis H_0.

Given that $F = \frac{s_1^2}{s_2^2} = 1.412636$ doesn't lie in the interval above, so we can't reject the null hypothesis, which means that the varicane of smoking and non-smoking mother are equal at 5% significance level.

```
# separate smoking group and non-smoking group
smoke <- filter(birthwt, smoke == 1)
non_smoke <- filter(birthwt, smoke == 0)
# F statistic
f <- var(smoke$lwt) / var(non_smoke$lwt)
# calculate two-sided 95% CI
df1 <- nrow(smoke) - 1
df2 <- nrow(non_smoke) - 1
c(qf(0.025, df1, df2), qf(0.975, df1, df2))
```

```
## [1] 0.6518345 1.5046602
```

```
# calculate F value
var(pull(smoke, lwt)) / var(pull(non_smoke, lwt))
```

```
## [1] 1.412636
```

b) Given your answer from part a, what kind of hypothesis test will you perform?

Two-sided T-test for testing the equality of mean in 2 samples with equal variance.

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.

Denote mean of smoking mother population is μ_1 and mean of non-smoking mother population is μ_2.

Test hypothesis : $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$,then $t = \frac{\overline{X_1} - \overline{X_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} = t_{187}$.

According to the former analysis, at 10% significance level,, when $t < t_{187,0.05}$ or $t > t_{187,0.95}$, which means \$t < -1.653043 \$ or $t > 1.653043$, we reject the null hypothesis H_0.

Given that $t = -0.6047303$ doesn't lie in the interval above, so we can't reject the null hypothesis, which means that the smoking status is not related to weight.

```
n1 <- nrow(smoke)
n2 <- nrow(non_smoke)
pooled_s <- sqrt(((n1-1)*var(smoke$lwt) + (n2-1) * var(non_smoke$lwt)) / (n1+n2-2))
t <- (mean(smoke$lwt) - mean(non_smoke$lwt)) / (pooled_s * sqrt(1/n1 + 1/n2))
# calculate two-sided 90% CI
c(qt(0.05, n1+n2-2), qt(0.95, n1+n2-2))
```

```
## [1] -1.653043  1.653043
```

Problem 3

According to the CDC, approximately 20% of pregnant American women suffer from hypertension. 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?

According to this scenario, we denote the distribution of the proportion of pregnant American women suffering from hypertension is X, then $\overline{X} \sim N(p, \frac{p(1-p)}{n}) = N(0.0634, 0.0003)$.

So the 99% confidence interval of $\overline{X}$ is $(\hat{p} - z_{0.995}\sqrt{\frac{p(1-p)}{n}}, \hat{\mu} + z_{0.995}\sqrt{\frac{p(1-p)}{n}})$, which equals to $(0.01768277, 0.10930136)$.

Interpretation: We are 99% confident that the true proportion of pregnant American women suffering from hypertension lies between 0.01768277 and 0.10930136.

Conclusion: Given that the 0.2 doesn't lie in the interval $(0.01768277, 0.10930136)$, so we are 99% confident that the statement that 20% of pregnant American women suffeingr from hypertension is invalid according to `birthwt` dataset.

```
# mean
p = mean(pull(birthwt, ht))
# standard error
se = sqrt((1 - p) * p / nrow(birthwt))
# two-sided 99% CI
c(p - qnorm(0.995) * se, p + qnorm(0.995) * se)
```

```
## [1] 0.01780412 0.10918001
```

b) Conduct a one-sided hypothesis test at the $\alpha = 0.1$ 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?

According to the scenario and former analysis, we test hypothesis: $H_0 : p = 0.2$ vs $H_1 : p < 0.2$

Given that $\overline{X} \sim N(p, \frac{p(1-p)}{n})$, then we could reject the null hypothesis when $Z = \frac{\overline{X}-0.2}{\sqrt{\frac{p(1-p)}{n}}}, Z < z_{0.1}$, which equals to $Z < -1.281552$ at 0.1 significance level.

Because $Z = -4.691685 < -1.281552$, so we could reject the null hypothesis. So we can conclude that the true proportion is indeed less than 20% in significance level 0.1, which means that the CDC's claim is NOT convincing at the $\alpha = 0.1$ level.

```
hypo_p = 0.2
se = sqrt((1 - hypo_p) * hypo_p / nrow(birthwt))
# calculate Z-score
z <- (p-hypo_p) / se
qnorm(0.1)
```

```
## [1] -1.281552
```

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 $\alpha = 0.01$ level. What can we conclude about the proportions of women with uterine irritability between the smoking groups?

Denote the proportions of women with uterine irritability in smoking mother population is p_1 and the proportions of women with uterine irritability in non-smoking mother population is p_2, the proportions of women with uterine irritability in smoking mother sample is $\hat{p_1}$ and the proportions of women with uterine irritability in non-smoking mother sample is $\hat{p_2}$.

Because $n_1p_1 > 10, n_1(1-p_1) > 10, n_2p_2 > 10, n_2(1-p_2) > 10$, so we can use normal approximation to estimate this binomial distribution.

Use two-sample binomial test for proportion.

Test Hypothesis: $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$ at significance level 0.01.

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

Rejection condition: $z < z_{0.005}$ or $z > z_{0.995} \Rightarrow z < -2.575829$ or $z > 2.575829$

Given that $z = 0.8545449$ doesn't lie in rejection interval, so we couldn't reject null hypothesis, which means the variance of the proportion of women with uterine irritability between non-smoking mother and smoking mother is equal at significance level 0.01.

```
# estimate proportion
p1 <- mean(pull(smoke, ui))
p2 <- mean(pull(non_smoke, ui))
# sample size
n1 <- nrow(smoke)
n2 <- nrow(non_smoke)
# calculate z-score
p <- (n1 * p1 + n2 * p2) / (n1 + n2)
z <- (p1-p2)/sqrt(p * (1-p) * (1/n1 + 1/n2))
# two-sided 99% interval
c(qnorm(0.005), qnorm(0.995))
```

```
## [1] -2.575829  2.575829
```

Conclusion: The proportion of women with uterine irritability between non-smoking mother and smoking mother are equal at significance level 0.01.

Problem 5

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

a) What test would be most appropriate to answer this question?

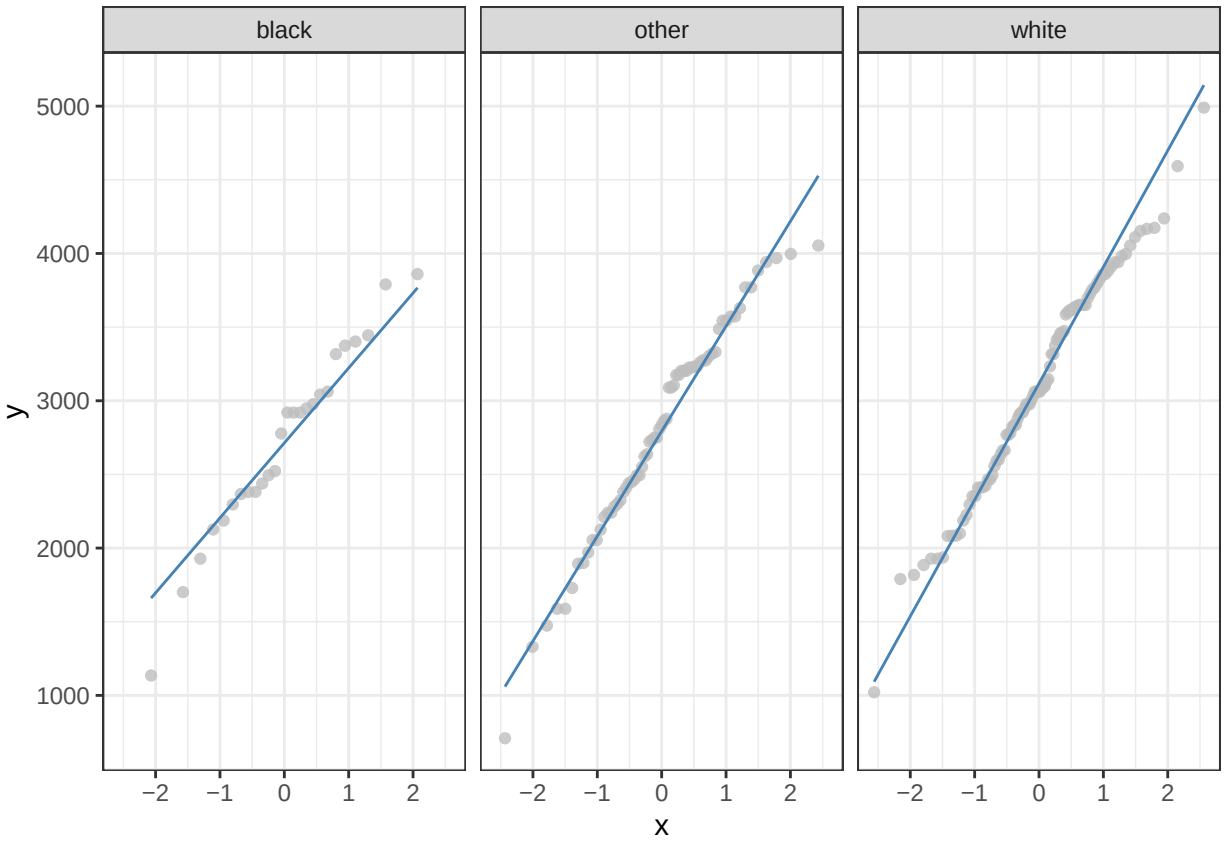
One-way ANOVA to test mean difference of 3 sample.

Supplement: Two-sided F-test to test the variance equality.

b) What assumptions are we making by using this test? Are all assumptions met?

1. There are more than 2 populations of interest: met, 3 population
2. The samples are drawn independently from the underlying populations: met
3. The variances of the k populations are equal: met, the hypothesis test process is described in the next.
4. The distributions of the error terms are normal: met (according to QQ-plot result)

```
birthwt |>
  mutate(race = recode(race,
                       `1` = "white",
                       `2` = "black",
                       `3` = "other")) |>
  ggplot(aes(sample = bwt)) +
  geom_qq(alpha = .8, color = "grey") +
  geom_qq_line(color = "steelblue") +
  theme_bw() +
  facet_grid(. ~ race)
```



Test variance equality:

Because it is a multiple comparison testing, use Bonferroni adjustment to control error rate.

For race i, denote the variance of birth weight of race i population is σ_i^2, and the variance of birth weight of race i sample is s_i^2.

Test hypothesis for race i and j: $H_0 : \sigma_i^2 = \sigma_j^2$ vs $H_1 : \sigma_i^2 \neq \sigma_j^2$ at significance level $\frac{0.05}{3}$.

$F = \frac{s_i^2}{s_j^2} \sim F_{n_i-1,n_j-1}$

Rejection condition: $F < F_{n_i-1,n_j-1,0.025}$ or $F > F_{n_i-1,n_j-1,0.975}$

1. "white" race and "black" race:

   Rejection condition: $F < 0.4965392$ or $F > 2.3549463$

   Conclusion: $F = 1.298838$, NOT in rejection condition, so accept H_0 for "white" race and "black" race at significance level $\frac{0.05}{3}$

2. "white" race and "other" race:

   Rejection condition: $F < 0.5853056$ or $F > 1.7503674$

   Conclusion: $F = 1.015825$, NOT in rejection condition, so accept H_0 for "white" race and "other" race at significance level $\frac{0.05}{3}$

3. "black" race and "other" race:

   Rejection condition: $F < 0.4149861$ or $F > 2.1111931$

   Conclusion: $F = 0.7821026$, NOT in rejection condition, so accept H_0 for "black" rance and "other" race at significance level $\frac{0.05}{3}$

Final Conclusion: The variance of birth weight of 3 races are equal.

```r
# two-sided F-test for variance equality of 2 samples
test_race_variance <- function(data, race_id1, race_id2, alpha){
  sample1 <- filter(data, race == race_id1)
  sample2 <- filter(data, race == race_id2)
  f <- var(pull(sample1, bwt)) / var(pull(sample2, bwt))
  df1 <- nrow(sample1) - 1
  df2 <- nrow(sample2) - 1
  result <- list()
  result$f_score <- f
  result$ci <- c(qf(alpha/2 , df1, df2), qf(1-alpha/2, df1, df2))
  result$alpha <- alpha
  return(result)
}
```

```r
# 1 = white, 2 = black, 3 = other
# use bonferroni adjustment
alpha = 0.05 / choose(3, 2)
# white vs. black
test_race_variance(birthwt, 1, 2, alpha)
```

```
## $f_score
## [1] 1.298838
##
## $ci
## [1] 0.4965392 2.3549463
##
## $alpha
## [1] 0.01666667
```

```r
# white vs. other
test_race_variance(birthwt, 1, 3, alpha)
```

```
## $f_score
## [1] 1.015825
##
## $ci
## [1] 0.5853056 1.7503674
##
## $alpha
## [1] 0.01666667
```

```r
# black vs. other
test_race_variance(birthwt, 2, 3, alpha)
```

```
## $f_score
## [1] 0.7821026
##
## $ci
## [1] 0.4149861 2.1111931
##
## $alpha
## [1] 0.01666667
```

c) Conduct the test at the 5% significance level and interpret your results. Be sure to write the hypotheses you are testing.

For race i, denote the mean of birth weight of race i population is μ_i^2.

Test hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$ vs H_1 : at least 2 population mean differ at significance level 0.05.

$F = \frac{\text{Between SS}/2}{\text{Within SS}/187} \sim F_{2,187}$

Rejection condition: $F > F_{2,186,0.95}$, which is $F > 3.044504$.

Given that $F = 4.912513 > 3.044504$ lies in rejection interval, so we could reject null hypothesis, which means there are at least 2 race population with different mean birth weight at significance level 0.01. So the race is related to birth weight.

```
# ANOVA
k <- 3
n <- nrow(birthwt)
y2n<- sum(pull(birthwt, bwt))^2 / n
sample_size <- c(nrow(filter(birthwt, race == 1)),
                 nrow(filter(birthwt, race == 2)),
                 nrow(filter(birthwt, race == 3)))
sample_mean <- c(mean(pull(filter(birthwt, race == 1), bwt)),
                 mean(pull(filter(birthwt, race == 2), bwt)),
                 mean(pull(filter(birthwt, race == 3), bwt)))
total_ss <- sum(pull(birthwt, bwt)^2) - y2n
between_ss <- sum(sample_size * sample_mean^2) - y2n
within_ss <- total_ss - between_ss
# calculate F-score
f <- (between_ss / (k-1)) / (within_ss / (n-k))
# critical value
qf(0.95, k-1, n-k)
```

```
## [1] 3.044504
```

d) Perform multiple comparisons - which races are significantly different? Interpret your results.

For race i, denote the mean of birth weight of race i population is μ_i^2.

Because it is a multiple comparison testing, use Bonferroni adjustment to control error rate.

Test hypothesis for race i and j: $H_0 : \mu_i = \mu_j$ vs $H_1 : \mu_i \neq \mu_j$ at significance level 0.05.

$t = \frac{\hat{\mu_i} - \hat{\mu_j}}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{n_i+n_j-2}$

Rejection condition: $t < t_{n_i+n_j-2,0.025}$ or $t > t_{n_i+n_j-2,0.975}$

1. "white" race and "black" race:
   Rejection condition: $t < -2.428004$ or $t > 2.428004$
   Conclusion: $t = 2.43935$, in rejection condition, so reject H_0 for "white" race and "black" race at significance level $\frac{0.05}{3}$
2. "white" race and "other" race:
   Rejection condition: $t < -2.419253$ or $t > 2.419253$
   Conclusion: $t = 2.57513$, in rejection condition, so reject H_0 for "white" race and "other" race at significance level $\frac{0.05}{3}$
3. "black" race and "other" race:
   Rejection condition: $t < -2.43904$ or $t > 2.43904$
   Conclusion: $t = -0.5290079$, NOT in rejection condition, so reject H_0 for "black" race and "other" race at significance level $\frac{0.05}{3}$

Final Conclusion: The "white" race is different from "black" and "other" races.

```
# two-sided T-test for mean equality testing of 2 sample with equal variance
test_race_mean <- function(data, race_id1, race_id2, alpha){
  sample1 <- filter(data, race == race_id1)
  sample2 <- filter(data, race == race_id2)
  df1 <- nrow(sample1) - 1
  df2 <- nrow(sample2) - 1
  # pooled sample standard deviation
  s <- sqrt((df1 * var(pull(sample1, bwt)) + df2 * var(pull(sample2, bwt))) / (df1 + df2))
  # T-score
  t <- (mean(pull(sample1, bwt)) - mean(pull(sample2, bwt))) / (s * sqrt(1 / nrow(sample1) + 1 / nrow(sa

  result <- list()
  result$t_score <- t
  result$ci <- c(qt(alpha/2, df1+df2), qt(1-alpha/2, df1+df2))
  result$alpha <- alpha
  return(result)
}
```

```
# white vs. black
alpha = 0.05 / choose(3,2)
test_race_mean(birthwt, 1, 2, alpha)
```

```
## $t_score
## [1] 2.43935
##
## $ci
## [1] -2.428004  2.428004
##
## $alpha
## [1] 0.01666667
```

```
# white vs. other
test_race_mean(birthwt, 1, 3, alpha)
```

```
## $t_score
## [1] 2.57513
##
## $ci
## [1] -2.419253  2.419253
##
## $alpha
## [1] 0.01666667
```

```
# black vs. other
test_race_mean(birthwt, 2, 3, alpha)
```

```
## $t_score
## [1] -0.5290079
##
## $ci
```

```
## [1] -2.43904  2.43904
##
## $alpha
## [1] 0.01666667
```