kx2224_hw3

Kx2224 Kangyu Xu

2024-11-02

Import packages

```r
library(MASS)
```

Import Original data

```r
head(birthwt)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

Problem 1

Construct a 95% confidence interval

```r
lwt_conf_interval = t.test(birthwt$lwt, conf.level = 0.95)$conf.int
lwt_conf_interval
```

```
## [1] 125.4270 134.2027
## attr(,"conf.level")
## [1] 0.95
```

Interpretion

The 95% confidence interval [125.4269766, 134.2026531] indicates that, if we were to take many samples, we expect the average pre-pregnancy weight of mothers to fall within this range 95% of the time. This interval provides an estimate of the true mean weight for the population.

Comments

1. Different Populations: The confidence interval we calculated is for a specific group—women who are pregnant, and it represents their pre-pregnancy weight. So if the statement is valid, it means that whether have pregnanted influences quite significantly on the distribution of women's weight.

2. Significant Difference in Values: The average weight for all American women is claimed to be 171 pounds, which is well outside our confidence interval of [125.43, 134.20] for pre-pregnancy weight. This large discrepancy suggests that the average pre-pregnancy weight of mothers is significantly lower than the average weight of the general female population.

3. Limitations of Direct Comparison: Since our data only reflects women who have been pregnant, it is not directly comparable to a general population statistic. Pregnant or previously pregnant women may differ in weight due to health factors, age distribution, or lifestyle compared to women who have never been pregnant.

Problem 2

F Test

To test for the equality of variances between the two groups (smoking and non-smoking mothers), we can use an F-test. The null hypothesis is that the variances are equal.

```r
# Separate the weights by smoking status
smoking_weights = birthwt$lwt[birthwt$smoke == 1]
nonsmoking_weights = birthwt$lwt[birthwt$smoke == 0]

# Conduct F-test for equality of variances
var_test_result = var.test(smoking_weights, nonsmoking_weights)
var_test_result
```

```
##
##  F test to compare two variances
##
## data:  smoking_weights and nonsmoking_weights
## F = 1.4126, num df = 73, denom df = 114, p-value = 0.09744
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9388406 2.1671700
## sample estimates:
## ratio of variances
##           1.412636
```

As the p-value $0.0974397 > 0.05$, so we cannot reject the null hypothesis. So these variances between the two groups are equal.

Hypothesis Test

Based on the result of the F-test, as variances are equal, I will use a two-sample t-test with equal variances.

Conduction

- H_0 : smoking status is related to weight
- H_1 : smoking status is not related to weight

```
t_test_result = t.test(smoking_weights, nonsmoking_weights, var.equal = TRUE, conf.level = 0.90)
t_test_result
```

```
##
##  Two Sample t-test
##
## data:  smoking_weights and nonsmoking_weights
## t = -0.60473, df = 187, p-value = 0.5461
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -10.306448   4.785414
## sample estimates:
## mean of x mean of y
##  128.1351  130.8957
```

As the p-value (0.5460913) is greater than 0.10, we fail to reject the null hypothesis, suggesting no significant difference. Based on the result, you would interpret whether there is evidence to support the claim that smoking status is related to pre-pregnancy weight.

Problem 3

Conduct a 99% confidence interval

```
hypertension_prop = mean(birthwt$ht == 1)
n = nrow(birthwt)
z = qnorm(0.995)  # z-score for 99% confidence

# Confidence interval calculation
margin_error = z * sqrt((hypertension_prop * (1 - hypertension_prop)) / n)
ci_lower = hypertension_prop - margin_error
ci_upper = hypertension_prop + margin_error
cat("99% Confidence Interval:", ci_lower, "-", ci_upper, "\n")
```

```
## 99% Confidence Interval: 0.01780412 - 0.10918
```

- **Interpretation**: The 99% confidence interval suggests that we are 99% confident that the true proportion of pregnant women with hypertension in our sample lies between 1.78% and 10.92%.
- **Conclusion**: Since 20% falls outside of our confidence interval, our data does not support the CDC's claim. Instead, our data suggests that the proportion of pregnant women with hypertension is likely much lower than 20%.

Hypothesis test

- H_0 : true proportion is indeed not less than the claimed 20%
- H_1 : true proportion is indeed less than the claimed 20%

```
# b) One-sided hypothesis test
p_null = 0.20
z_score = (hypertension_prop - p_null) / sqrt((p_null * (1 - p_null)) / n)
p_value = pnorm(z_score)
cat("P-value for the one-sided test:", p_value, "\n")
```

```
## P-value for the one-sided test: 1.354823e-06
```

```
# Interpretation
if (p_value < 0.1) {
  cat("We reject the null hypothesis and conclude that the true proportion is less than 20%.\n")
} else {
  cat("We do not reject the null hypothesis; there is insufficient evidence to conclude that the true p
}
```

```
## We reject the null hypothesis and conclude that the true proportion is less than 20%.
```

As the p_value is smaller than 0.1, we reject the null hypothesis and conclude that the true proportion is less than 20%.

Problem 4

- H_0 : There is no difference in the proportions of uterine irritability between smokers and non-smokers.
- H_1 : There is a difference in the proportions of uterine irritability between smokers and non-smokers.

```
smoke_group = birthwt$ui[birthwt$smoke == 1]
non_smoke_group = birthwt$ui[birthwt$smoke == 0]

p1 = mean(smoke_group)  # Proportion of uterine irritability in smokers
p2 = mean(non_smoke_group)  # Proportion of uterine irritability in non-smokers
n1 = length(smoke_group)
n2 = length(non_smoke_group)

prop.test(x = c(sum(smoke_group), sum(non_smoke_group)),
          n = c(n1, n2),
          alternative = "two.sided",
          conf.level = 0.99)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(smoke_group), sum(non_smoke_group)) out of c(n1, n2)
## X-squared = 0.41576, df = 1, p-value = 0.5191
## alternative hypothesis: two.sided
```

```
## 99 percent confidence interval:
##  -0.1056061  0.1960879
## sample estimates:
##    prop 1    prop 2
## 0.1756757 0.1304348
```

I used a two-proportion z-test to compare the proportions at a significance level of $\alpha = 0.05$. As the p-value is greater than or equal to 0.01, we fail to reject the null hypothesis, suggesting no statistically significant difference in uterine irritability between the two groups.

Problem 5

Choose hypothesis test method

Since we want to see if there is a difference in birth weights (`bwt`) across different racial groups (`race`), the most appropriate test would be a one-way ANOVA (Analysis of Variance). This test allows us to compare the means of birth weights among multiple groups (races) to see if there is a statistically significant difference.

Assumption

For ANOVA to be valid, we need to make the following assumptions:

- Independence: The samples are independent of each other.
- Normality: The birth weights in each racial group are approximately normally distributed.
- Homogeneity of Variance: The variances in birth weight across racial groups are approximately equal.

Yes, all the assumptions met.

Conducting the ANOVA Test

- H_0 : The mean birth weight is the same across racial groups.
- H_1 : At least one racial group has a different mean birth weight.

```r
birthwt$race = as.factor(birthwt$race)

# Conduct one-way ANOVA
anova_result = aov(bwt ~ race, data = birthwt)
summary(anova_result)
```

```
##              Df   Sum Sq Mean Sq F value  Pr(>F)   
## race          2  5015725 2507863   4.913 0.00834 **
## Residuals   186 94953931  510505                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p-value from the ANOVA test is less than 0.05, we reject the null hypothesis, indicating a significant difference in birth weight among racial groups.

Multiple comparison

```r
# Perform Tukey's HSD test for multiple comparisons
tukey_result = TukeyHSD(anova_result)
print(tukey_result)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = bwt ~ race, data = birthwt)
##
## $race
##          diff       lwr        upr     p adj
## 2-1 -383.02644 -756.2363  -9.816581 0.0428037
## 3-1 -297.43517 -566.1652 -28.705095 0.0260124
## 3-2   85.59127 -304.4521 475.634630 0.8624372
```

The Tukey test results will show which racial groups have statistically significant differences in birth weight. As the p-value of 2-1 & 3-1 groups are less than 0.05, so the group 2 and 3 have significant differences in brith weight with group 1.