HW4: Data Analysis of Zinc Concentration in Two Water Sources

Yunchun Pan, Siqing Lyu, Nathan Ng, Pudan Xu

2/2/2021

Introduction

Trace metals in drinking water affect the flavor, and an unusually high concentration can pose a health hazard. In this homework, we analyze the quality of drinking water dataset **water.txt**[1]. We closely examine zinc concentration in bottom water and in surface water, and we run statistical tests to check if there is a difference in population means of zinc concentration in two water sources. This water dataset contains 10 pairs of data and has the following two variables:

- **bottom**: the zinc concentration in bottom water
- **surface**: the zinc concentration in surface water

Under the assumption that each zinc concentration in bottom water is uniquely paired to a zinc concentration in surface water, we first perform one sample t-test and an univariate Hotelling's test for paired data. Then, supposing they are not uniquely paired, we perform a two-sample Hotelling's test. In the end, we compare the results of tests for paired and unpaired data and check if they lead us to different conclusions.

Our Work

We first load the drinking water dataset into R to get the data we use throughout this assignment.

```
water <- read.delim("../Data/Water.txt")
water
```

```
##    bottom surface
## 1   0.430   0.415
## 2   0.266   0.238
## 3   0.567   0.390
## 4   0.531   0.410
## 5   0.707   0.605
## 6   0.716   0.609
## 7   0.651   0.632
## 8   0.589   0.523
## 9   0.469   0.411
## 10  0.723   0.612
```

```
col_names <- c(colnames(water))
water_means <- c(mean(water$bottom), mean(water$surface))
water_sd <- c(sd(water$bottom), sd(water$surface))
explore_data <- data.frame("Variable"=col_names,
                           "Sample_Mean"=water_means,
                           "Sample_Std"=water_sd)
explore_data
```

[1]Source: Dataset provided in the Math 189 Github Data Repository: https://github.com/tuckermcelroy/ma189/blob/main/Data/Water.txt

```
##   Variable Sample_Mean Sample_Std
## 1   bottom      0.5649  0.1467814
## 2  surface      0.4845  0.1312294
```

The table above shows sample means and sample standard deviations of zinc concentration in bottom water and surface water separately. There is a higher average zinc concentration in bottom water of 0.5649, compared to average zinc concentration in surface water of 0.4845. Its concentration in both water sources seems to have a similar spread, although zinc concentration in bottom water has a slightly higher standard deviation of 0.1467814, compared to that in surface water, which has a standard deviation of 0.1312294.

We denote the population mean of zinc concentration in bottom water by μ_1 and the population mean of zinc concentration in surface water by μ_2. Then, we want to test the null and alternative hypotheses:

$$H_0 : \mu_1 = \mu_2 \qquad H_a : \mu_1 \neq \mu_2$$

Our null hypothesis states that there is no difference in population means of zinc concentration in bottom and surface water, while alternative hypothesis states there is a difference.

Assuming that each zinc concentration in bottom water is uniquely paired to a zinc concentration in surface water, We re-write the null and alternative hypotheses as such:

$$H_0 : \mu^{(1)} - \mu^{(2)} = 0 \quad H_a : \mu^{(1)} - \mu^{(2)} \neq 0$$

Here, We denote zinc concentration in ith bottom water sample as $x_i^{(1)}$ and that in ith surface water sample as $x_i^{(2)}$. Then, we denote y_i as the difference in zinc concentration of the ith pair (ith bottom water sample and ith surface water sample):

$$y_i = x^{(1)} - x^{(2)}$$

The population mean of y_i is denoted as u_y, which is the difference in population means of zinc concentration in botton water and surface water:

$$u_y = u^{(1)} - u^{(2)}$$

This way, we transform this two-sample testing problem to a one-sample problem:

$$H_0 : \mu_y = 0 \quad H_a : \mu_y \neq 0$$

First, we perform a one-sample t-test on our paired data. The sample mean is the difference between sample means of zinc concentration in bottom water and surface water. The null hypothesis of our test is that there is no difference in average zinc concentration in two water sources, expressed in notation as $\mu_y = 0$.

```
alpha <- 0.05
n <- dim(water)[1]
zinc_diff <- water[,1] - water[,2]
t_stat <- (mean(zinc_diff) - 0)/(sd(zinc_diff)/sqrt(n))
crit_value <- qt(1-alpha/2, df=n-1)
p_value <- 2*pt(abs(t_stat), df=n-1, lower.tail=FALSE)
t_test <- data.frame("T_Statistic"=t_stat,
                     "Critical_Value"=crit_value,
                     "P_Value"=p_value, "Significance_Level"=alpha)
t_test
```

```
##   T_Statistic Critical_Value      P_Value Significance_Level
## 1    4.863813       2.262157 0.0008911155               0.05
```

In the paired sample t-test, we calculate a t-statistic of 4.864, which is more than double the critical value at a 0.05 significance level. Moreover, the p-value of 0.00089 is much smaller than the significance level of 0.05. Because our test statistic is greater than the critical value and the p-value is smaller than significance level of 0.05, we reject the null hypothesis that the population means of zinc concentration in bottom water and surface water are equal.

Now, let's perform a Hotelling's T^2 test for paired data. We square up the t-statistic we calculate above to a F statistic and test on an F distribution with 1 and n-1 degrees of freedom.

```
m <- 1
f_stat <- t_stat^2
crit_value <- qf(1-alpha,df1=1,df2=n-1)
p_value <- pf(f_stat, df1=1, df2=n-1, lower.tail=FALSE)
f_test <- data.frame("F_Statistic"=f_stat,
                     "Critical_Value"=crit_value,
                     "P_Value"=p_value, "Significance_Level"=alpha)
f_test
```

```
##   F_Statistic Critical_Value      P_Value Significance_Level
## 1    23.65667       5.117355 0.0008911155               0.05
```

After performing a Hotelling's T^2 test for paired data, we find that the F-statistic is 23.65667, which is larger than the critical value of 5.117355 by approximately 4 times. The p-value is 0.000891, and it is much lower than the significance level of 0.05. This p-value in the Hotelling's T^2 test is also equal to the p-value we calculate in the one-sample t-test above.

Just like one-sample t-test done before, this univariate Hotelling's T^2 test gives us clear evidence to reject the null hypothesis that the population means of zinc concentration in bottom water and surface water are equal.

Supposing that zinc concentrations in bottom water and surface water are not uniquely paired, we now apply a two-sample Hotelling's test to examine our null hypothesis. Since this drinking water dataset is small and has only 10 samples, we cannot ignore the estimation error of estimating the population covariance matrix with the sample counterpart. Hence, we cannot assume variances are the same, and we should use different estimate given by

$$\mathbf{S}_T = n_1^{-1}\mathbf{S}^{(1)} + n_2^{-1}\mathbf{S}^{(2)}$$

and adjust our test statistics accordingly to get the result.

We first calculate sample variances of zinc concentration in bottom and surface water, respectively. The results show that their variances are very close, even though zinc concentration in bottom water has a slightly higher variance of 0.02154477, compared to that in surface water, which has a variance of 0.01722117.

```
col_names <- colnames(water)
var_bottom <- var(water[,1])
var_surface <- var(water[,2])
data.frame("Variable"=col_names,
           "Variance"=c(var_bottom,var_surface))
```

```
##   Variable   Variance
## 1   bottom 0.02154477
## 2  surface 0.01722117
```

Then we calculate the weighted-average of sample covariance matrices $S^{(1)}$ and $S^{(2)}$ by following the formula below:

$$\mathbf{S}_T = n_1^{-1}\mathbf{S}^{(1)} + n_2^{-1}\mathbf{S}^{(2)}$$

```
n1 <- dim(water)[1]
n2 <- dim(water)[1]
var_weighted <- n1^{-1} * var_bottom + n2^{-1} * var_surface
round(var_weighted,digits=6)
```

```
## [1] 0.003877
```

Next, we calculate the T^2 statistic **hotel** and F statistic **f_stat**. To find the critical value of this two-sample Hotelling's test, we calculate the degree of freedom ν according to the following complicated formula:

$$\frac{1}{\nu} = \sum_{k=1}^{2} \frac{1}{n_k - 1} \left(\frac{(\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})' \mathbf{S}_T^{(1)} \left(n_k^{-1} \mathbf{S}^{(k)} \right) \mathbf{S}_T^{-1} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})}{T^2} \right)^2$$

```
z <- water$bottom - water$surface
hotel <- t(mean(z)) %*% solve(var_weighted) %*% (mean(z))
m <- 1
alpha <- .05
f_stat <- (n1+n2-m-1)/(m*(n1+n2-2)) * hotel
inv_v <- 1/(n1-1) *(((t(mean(z)) %*% solve(var_weighted) %*% (n1^{-1} * var_bottom) %*%
                        solve(var_weighted) %*% (mean(z)))/hotel)^2 +
                      ((t(mean(z)) %*% solve(var_weighted) %*% (n2^{-1} * var_surface) %*%
                          solve(var_weighted) %*% (mean(z)))/hotel)^2)
v =1/inv_v
critical_val = qf(1-alpha,df1=m,df2=v)
data.frame("F_Statistic"=f_stat,
           "Critical_Value"=critical_val)
```

```
##   F_Statistic Critical_Value
## 1    1.667485       4.421755
```

After performing this two-sample Hotelling's T^2 test for unpaired data, we find that the F-statistic is 1.667485, which is less than the critical value of 4.421755. Therefore, we fail to reject the null hypothesis that the population means of zinc concentration in bottom water and surface water are the same.

In both one-sample t-test and univariate Hotelling's T^2 test for paired data, test statistics calculated are larger than the critical value by two times or more, and the same p-value of 0.0008911155 in both tests is much smaller than the significance level of 0.05. Therefore, when zinc concentration in bottom water is paired with one in surface water, we have clear evidences to reject the null hypothesis that the population means of zinc concentration in bottom water and surface water are equal. In contrast, when the data is unpaired, test statistic of 1.667485 in two-sample Hotelling's test is less than the critical value of 4.42175, and hence we fail to reject the null hypothesis that the population means of zinc concentrations in bottom water and surface water are equal.

Since zinc concentration in bottom water are highly correlated with that in surface water, every zinc concentration in bottom water should be paired uniquely to a zinc concentration in surface water. Therefore, two-sample Hotelling's test might not be appropriate for testing our null hypothesis and it gives a different result in the end.

Conclusion:

Given the quality of drinking water dataset **water.txt**, we first explore sample means and sample standard deviations of zinc concentrations in bottom water and surface water. We find that, there is a higher average zinc concentration in bottom water than average zinc concentration in surface water. Its concentration in both water sources seems to have a similar spread, although zinc concentration in bottom water has a slightly higher standard deviation than that in surface water.

Supposing the zinc concentration in bottom water is uniquely paired to a zinc concentration in surface water and using the paired sample test, we perform one-sample t-test and univariate Hotelling's T^2 test for paired data. Test statistics calculated are larger than critical value by two times or more, and p-value we get from two tests is much smaller than the significance level of 0.05. Hence, we reject the null hypothesis that the population means of zinc concentration in bottom water and surface water are equal.

Supposing the data is not paired, we apply two-sample Hotelling's test. Because the drinking water dataset is small and only has 10 samples, we cannot ignore the estimation error of estimating the population covariance

matrix with the sample counterpart. Therefore, we can not assume the variances are the same and we should adjust our test statistics accordingly. In the two-sample Hotelling's test, test statistic is less than the critical value. Hence, we fail to reject the null hypothesis that the population means of zinc concentration in bottom water and surface water are equal.

Given different results from tests for paired and unpaired data, we believe two-sample Hotelling's test might not be appropriate for testing our null hypothesis. This is because zinc concentration in bottom water might be highly correlated with that in surface water. So, every zinc concentration in bottom water should be paired uniquely to a zinc concentration in surface water, and we should perform tests for paired data.