Math 189: Homework 4

Alan Lui, Derek So, Xiangyu Wei

February 8, 2021

Introduction

Trace metals in drinking water affect the flavor, and an unusually high concentration can pose a health hazard. The water quality dataset (water.txt) contains ten pairs of data that measure zinc concentration in bottom water and surface water.

This study looks into different concentration levels of zinc in water to determine whether the water is safe to drink. Using paired sample t-test and (unpaired) two sample Hotelling's test, we compare water at surface and bottom level to determine whether the zinc concentration is consistent across the body of water, and discuss if there is a difference between the outcome of the two different methods.

Tasks & Analysis

Dataset

The water quality dataset consists of 10 samples, in two variables of bottom and surface water. Given no units, we assume that the quantity of zinc concentration is in milligrams per liter, although in the end the exact units are irrelevant to the data. No relevant source is found, so only datahub citation link is included here.

Source: https://github.com/tuckermcelroy/ma189/blob/main/Data/Water.txt at 2021-02-04 20:08:31 PST.

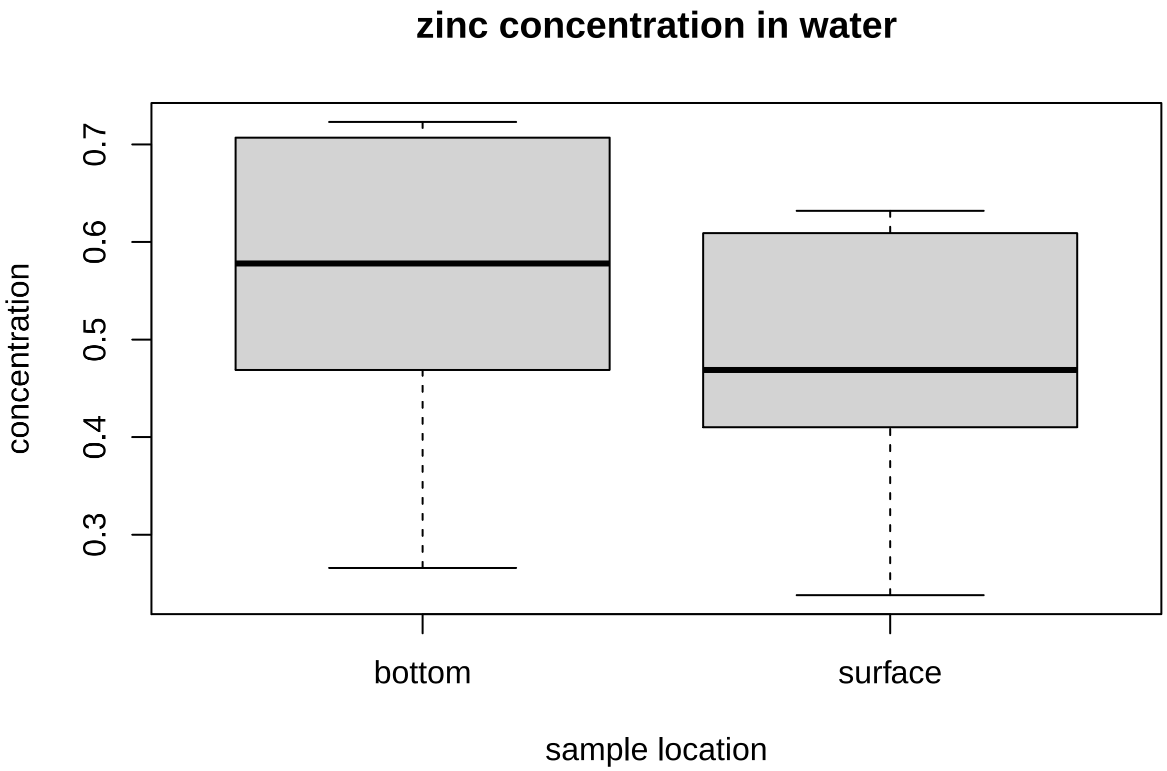
```r
#import dataset
Water <- read.table("Data/Water.txt", head = TRUE)
Water
```

```
##    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
```

Tasks

Let us first briefly explore the contents of our dataset to oversee any peculiarities:

```
boxplot(Water, xlab = "sample location", ylab = "concentration",
        main = "zinc concentration in water")
```



It is first important to note that there are very few samples to work with here (10) so there may be a higher variance with respect to the population. Looking at the boxplots, it shows that the median zinc concentration in bottom water (~0.59) is higher than that of surface water (~0.5). Also observe that zinc concentration in bottom water has a higher maximum value than that of surface water. Keeping these ideas in mind, we will begin with the analysis:

1. Suppose we consider the zinc concentration in bottom water and in surface water as two samples. Denote by μ_1 and μ_2 the underlying population means of the two samples. Test the null and alternative hypotheses
$$H_0 : \mu_1 = \mu_2 \qquad H_a : \mu_1 \neq \mu_2$$
using the paired sample test.

First, we apply the paired sample t-test [1] (which would yield the same result as using a paired Hotelling's T^2) to determine if there is a significant difference between the two sets of values. We are assuming that the samples obtained from the water are at roughly the same location with varying depth in order for the data to be accurately paired, and that the samples are independent of each other. Keep in mind that since the sample size is only 10, the conditions of the central limit theorem are not met. We will approach this test with caution.

```
x = Water[,1] #storing "bottom" zinc concentration
y = Water[,2] #storing "top" zinc concentration

# paired t test with paired = T
t.test(x, y, paired = T)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = 4.8638, df = 9, p-value = 0.0008911
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.043006 0.117794
## sample estimates:
## mean of the differences
##                  0.0804
```

The paired t-test gives the test statistic $t(9) = 4.8638, p = 0.0009$. Thus, with an $\alpha = 0.05$, we would reject the null hypothesis that the bottom and surface zinc concentrations are the same.

2. Suppose the data was not paired. Apply the two-sample Hotelling's test. Do you assume the variances are the same? Use the variance estimate that is appropriate for your decision, as to whether the variances are the same or different.

Then, as these values are dependent variables from the same body of water, we use the two-sample Hotelling's Test to check again. Since the sample size of 10 is critically small, we cannot assume that the variances are equal. Since we cannot use the assumption of equal variances, we will compute the degrees of freedom ν and will calculate the F transformation of the T^2 statistic [2].

```
# simplify calculations since n1 = n2 =10, and feature m = 1
m = 1
n = 10
S_T = (var(x) + var(y))/n
diff = mean(x) - mean(y)
T2 = diff^2/S_T
T2
```

```
## [1] 1.667485
```

```
# calculation for F
F = (2*n-m-1)/(m*(2*n-2)) * T2
F
```

```
## [1] 1.667485
```

```
# calculation for v
nu = 1 / ((((diff/S_T)^2*(var(x)/n))/T2)^2/(n-1) + (((diff/S_T)^2*(var(y)/n))/T2)^2/(n-1))
nu
```

```
## [1] 17.77885
```

The F stat follows a F distribution with degrees of freedom $(m = 1, \nu \approx 18)$ [2].

```r
pf(q = F, df1=m, df2=round(nu), lower.tail = FALSE)
```

```
## [1] 0.2129332
```

The two-sample Hotelling's T^2 gives the test statistic $F(1, 18) = 1.6675, p = 0.2129$. Thus, with an $\alpha = 0.05$, we would retain/fail to reject the null hypothesis that the bottom and surface zinc concentrations are the same.

Conclusion

3. Summarize how the results differ based on whether the samples are paired or unpaired.

Using the paired sample test, we can confidently reject the null hypothesis with a t of 4.86 and a p value of 0.000089 with 9 degrees of freedom using a significance level of $\alpha = .05$. The mean of the differences was 0.0804, which proportional to data points ranged from 10% to 33% difference from the base value.

Using the two sample Hotelling's test with unequal variances, we fail to reject the null hypothesis at a significance level of $\alpha = .05$ with a F statistic of 1.6675, a p-value of 0.2129, and degrees of freedom: m = 1 and $\nu = 18$ (rounded).

Note that we do have a discrepency between the results of the different tests we conducted on the samples. We will discuss this further in the conclusion below.

Discussion

The dataset itself is very limited and tells us nothing of the source, so we are rather limited in exploring possible implications. It is also important to consider the fact that the 10 sample size results in a high variance. Furthermore, a low sample size does not satisfy the condition of normality as larger samples, above 30 are preferred to invoke the central limit theorem.

We believe that since the samples of water were taken from the same body of water, the paired t-test would be the more appropriate test to consider. Using the paired t test, we reject the null hypothesis meaning that the differnece in concentration of zinc in the upper and bottom water samples are different.

Although the two sample Hotelling's test yielded a non-significant result, it is safer in this scenario to have a false positive over a false negative as the levels of zinc is harmful to the human body. Precautions taken from the results of the paired t test will only ever benefit the health of the population, bearing the cost of filtration.

Appendix

[1] Paired Sample Test:

The paired t test operations under the condition that the two samples obtained are naturally matched or paired, for example, medical trials regarding the same patient. It is also required that the samples be independent of one another, and be approximately normally distributed. The paired data are treated as one sample where the difference between the two values of interest are observed. The t statistic is calculated as follows:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

The t statistic follows a T distribution with n-1 degrees of freedom, where n is the sample size.

[2] 2-Sample Hotelling's Test (Small Sample)

The two sample Hotelling's Test operates on two samples and is a function of the difference between the sample means for the two populations. Due to small samples, we are unable to assume that the variances are equal. As a result, we will compute the statistic and degrees of freedom differently than that of a large sample.

F statistic transformation:

$$F = \frac{n_1 + n_2 - m - 1}{m(n_1 + n_2 - 2)} T^2 \sim \mathcal{F}_{m,\nu},$$

Degrees of freedom ν:

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

At significance level α, we reject null hypothesis H_0 if

$$F > \mathcal{F}_{m,\nu,\alpha}.$$