ANOVA, Model Assumption and Statistical Analysis on Effect of Household Processing Method to pH Value of Water

STA305 Final Report (LEC 0101), Group 22

Jiaqi Bi, 1003886609 Lei Cao, 1005715111 Lanruo Li, 1005149581
Le Shen, 1005935106 Yuika Cho, 1003213186

Contents

Introduction

The increased concern on water quality has been rigorously demanding, and this has been a public health issue for years. Primarily there are miscellaneous scientific articles and news about the pH value of water affects human health. The question of interest lies in if the household water processing method will affect the pH value. The experiment uses the most common approach that everyone can practice at home. After determining the sample size and the collection of experiment data, one-way ANOVA, linear regression modelling with dummy coding method, and contrasts analysis were practiced to check the statistical significance of the assumption. The main purpose of the project is to find the best way of producing an ideally higher pH value.

Experimental Design and Process

The household water processing methods include Stilled Water, Boiled Water, Filtered Water, and Frozen Water. The water source is controlled to be tap water as gaining water from the tap is widely habitual everywhere. Controlling the water source prevents excessive biases caused by different unforeseeable water purity. Moreover, the initial volume of the cup of water is controlled. Indeed, the experimental unit is the cup of water. To reduce the confoundings of the unit, a Between-Subjects design is practiced throughout. As mentioned, there are 4 levels of the predictor variable. Due to the filling time of the water may be different and hard to be controlled or blocked, using randomization that after filling all cups of water by labeling each cup randomly is the best way to minimize the "time effect" of output. The repliaction is easy to follow that the water source could be selected otherwise. There are total 48 cups of water with each has a unique number, such that label 1-12 cups are given the treatment of being frozen, 13-24 to be filtered, 25-36 to be boiled, and 37-48 to be stilled (Control Group).

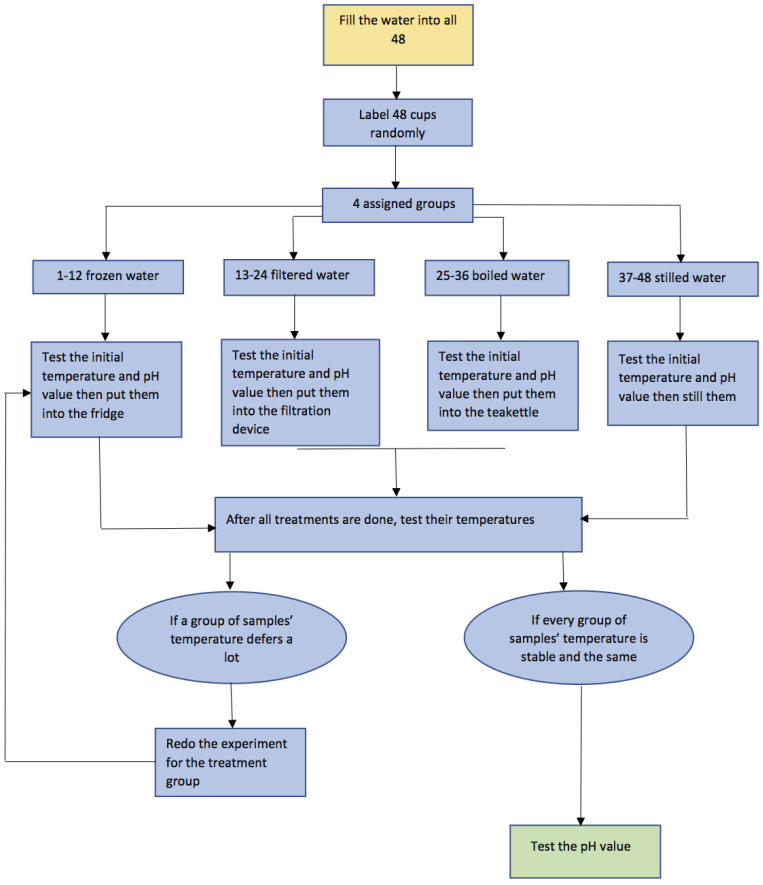
Figure 1: Experiment Conduction with pH Meter and Thermometer

The stilled water treatment can be treated as a control group to avoid the time being a factor of the pH value change. Blindness is not essential for this experiment as the output of pH value is examined by a specific meter (objective measurement). There is another meter involved in the experiment: the thermometer to ensure each treatment level has the same initial temperature, and each unit of a level has the same temperature after the treatment. Figure 1 shows these two testers and the experiment conduction.

Specific treatment details:

- Stilled Water: Control Group, stilling the group until the end of the experiment.
- Filtered Water: Using the same water filtration device to filter all cups of this group.
- Frozen Water: Using the same fridge to freeze all cups of this group simultaneously.
- Boiled Water: Using the same teakettle to boil all cups of this group simultaneously..

Specific experiment flowchart:



Data Analysis

Sample Determination

Before the conduction of the experiment, the determination of sample size is finalized by using Balanced one way ANOVA power calculation: Manually set the power to be $f = 0.8$ for a large effect, that the result of sample size determination turns out to be $n \doteq 12$.

Data Visualization

Figure 2 is a general frame of the collected data where the data visualization is practiced. Instead of showing the pH value output, the graph illustrates the pH value difference of before and after experiment for each treatment.

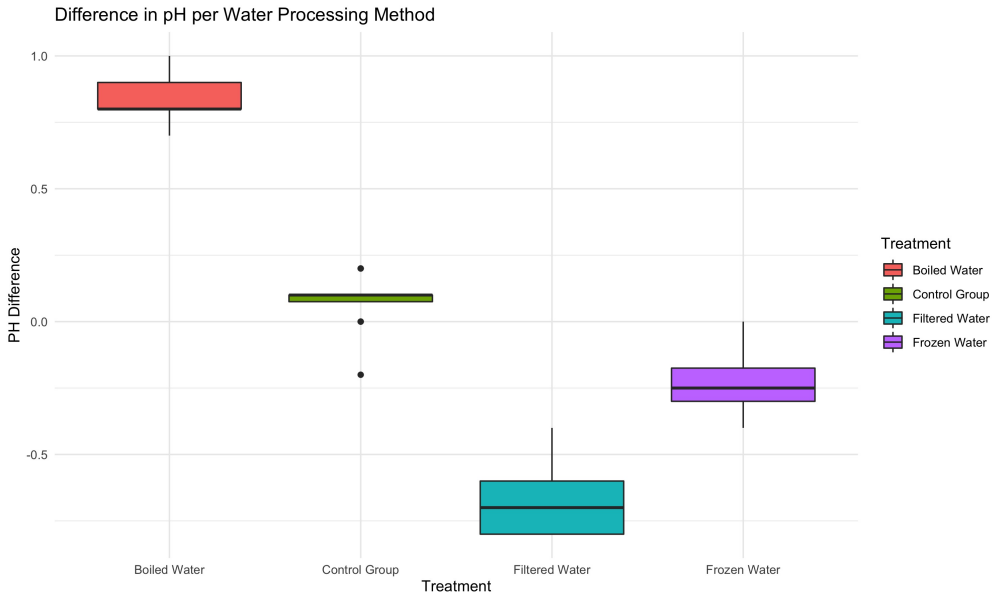


Figure 2: Boxplot of Difference in pH per Water Processing Method

Mean Value and Statistical Modeling

While a boxplot shows intuitionistic results, the statistical significance of the result needs a deeper analysis. The mean value and variance of each treatment show that:

Table 1: Table of Mean Value and Variance of Data

Treatment	Mean Value	Variance
Bolied Water	$+0.842$	0.0081
Filtered Water	-0.675	0.0166
Frozen Water	-0.225	0.0130
Stilled Water	$+0.075$	0.0111

The linear regression of the data using dummy coding scheme:

$$\hat{Y}_i = 0.0750 + 0.7667I_{Boiled,i} - 0.7500I_{Filtered,i} - 0.3000I_{Frozen,i}$$

For $m = Boiled, Filtered, Frozen$ as Stilled Water is the default and has no indicators,

$$I_{m,i} = \begin{cases} 1, & \text{if } i\text{th case is in level } m \\ 0, & \text{otherwise} \end{cases}$$

One-Way ANOVA and Testing Contrasts

The null hypothesis is all mean pH value differences of 4 water processing methods are the same. The alternative hypothesis is at least one pair of mean pH value differences are not the same. The result shows a significance (p-value < 0.05) of the F-test. That is, the null hypothesis is rejected. Thus, there is a strong evidence to conclude that different water processing methods defer in pH value difference.

The formula of linear contrast is written as a linear combination of the group means (mean difference in pH value):

$$\psi = \sum_{j=1}^{4} c_j \mu_j$$

where j indicates each level, c_j is the coefficients in the contrast that needs to be restricted as the sum of coefficients to be 0:

$$\sum_{j=1}^{4} c_j = 0$$

The experiment contains 4 types of water processing, and it has 3 temperature levels after assigning treatment to those samples. The temperature level is marked as "Medium" when the temperature after the experiment is greater than 15 and less than 45, "Low" when the temperature after the experiment is less than 15, "High" when the temperature afterr the experiment is greater or equal to 45. Table 2 shows everry linear combination of group means and all comparisons.

Table 2: Table of All Comparisons and P-value

Group	Stilled	Boiled	Filtered	Frozen	Contrast Estimate	p-value
Temp Level	Medium	High	Medium	Low		
Means ($\bar{y_j}$)	0.075	0.8417	−0.6750	−0.2250	$\hat{\psi}$	
ψ_1	1	−1	0	0	−0.7667	$< 2 \times 10^{16}$
ψ_2	0	1	−1	0	1.5167	$< 2 \times 10^{16}$
ψ_3	0	0	1	−1	−0.4500	7.09×10^{13}
ψ_4	1	−2	1	0	−2.2834	$< 2 \times 10^{16}$
ψ_5	1	0	1	−2	−0.1500	0.0613
ψ_6	0	1	0	−1	1.0667	$< 2 \times 10^{16}$

Type I Error Solution

There are total 6 tests with $\alpha = 0.05$:

$$\alpha_{FW} = P(\text{at least one Type I Error}) = 1 - (1 - 0.05)^6 = 0.2649$$

The probability of having a Type I Error somewhere during tests is 26.49%. The test focuses on the pairwise comparison, and the experiment design is balanced. Using Tukey's HSD correction is the best way to solve high Type I error issue. There are $\binom{4}{2} = 6$ pairs of comparison in the Tukey's HSD. Table 3 shows the 95% family-wise confidence interval level.

Table 3: Tukey Multiple Comparisons of Means with 95% Family-Wise CI

Treatment Comparisons	Confidence Interval	$\hat{\psi}$
Boiled vs. Stilled	$(0.6463, 0.8870)$	0.7667
Filtered vs. Stilled	$(-0.8704, -0.6296)$	-0.7500
Frozen vs. Stilled	$(-0.4204, -0.1796)$	-0.3000
Filtered vs. Boiled	$(-1.6370, -1.3963)$	-1.5167
Frozen vs. Boiled	$(-1.1870, -0.9463)$	-1.0667
Frozen vs. Filtered	$(0.3296, 0.5704)$	0.4500

Conclusion

Based on the complex comparison, there is strong evidence that Boiled Water increases the pH value comparing to Stilled Water and Filtered Water. There is also strong evidence that Filtered Water decreases pH value comparing to Frozen Water, Higher temperature method increases pH value comparing to Low and Medium temperature level. However, the statistics show there is no evidence that Medium temperature decrease pH value comparing to Low temperature. Furthermore, the Tukey's CI reduces the Type I Error rate and gives a significant results of pairwise comparisons since all comparisons do not contain 0 value of CI. Thus, there is a strong evidence to deduce that different water processing techniques do change the pH values accordingly. The best way to increase pH value includes boiling the water with higher temperatures, and the best way to decrease pH value is to filtering the water while lowering the temperature.

Limitations and Discussion

The experiment only takes place with the object water, other substances are not considered. The temperature indeed plays a strong role in altering the pH value. There are more methods to consider the temperature as a factor, such as ANCOVA, or other variable selection criteria that may be better statistical methods to deal with quantitative factors. The experiment could have done better if there is one more water source to be chosen from. In this manner, Two-Way ANOVA can be developed and it expands the experiment details and diversity.

Appendix

The appendix contains all R codes and outputs that used during the experiment.

```
library(pwr)# Packages used in the data analysis
library(tidyverse)# Packages used in the data analysis
library(tidyr)# Packages used in the data analysis
library(ggplot2)# Packages used in the data analysis
library(multcomp)# Packages used in the data analysis
library(readxl)# Packages used in the data analysis
```

```
pwr.anova.test(k=4, n=NULL, f=0.5, sig.level=0.05, power=0.8)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 4
##               n = 11.92611
##               f = 0.5
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

```
#Sample Size Determination
```

```
Final.Data <- read_excel("STA305 FINAL PROJECT DATA.xlsx") #Read excel into R
Final.Data <- Final.Data[-c(49:55),]
#Delete unnecessary rows
Final.Data <- Final.Data %>%
  rename(
    pH.Before=`pH Before Treatment (0-14)`,
    pH.After=`pH After Treatment (0-14)`,
    Temp.Before=`Initial Temperature (℃)`,
    Temp.After=`Temperature After Treatment (℃)`
  ) %>%
  mutate(Final.Data,Temp.After.Level=
      ifelse(Temp.After>=15 &Temp.After<45,"Medium",
      ifelse(Temp.After<15, "Low",
      ifelse(Temp.After>=45, "High", "NA"))))
#Data Wrangling, to record Low Medium High Temperatures into dataset
Final.Data$DiffinPH=Final.Data$pH.After-Final.Data$pH.Before
#Calculate pH differences
attach(Final.Data)
#Attach the dataset
DiffinPH.1<-pH.After-pH.Before
#Calculate pH differences
```

```
DiffinPH.Frozen <- DiffinPH[Treatment=="Frozen Water"]
#Calculate pH differences for Frozen Water
DiffinPH.Filtered <- DiffinPH[Treatment=="Filtered Water"]
#Calculate pH differences for Filtered Water
DiffinPH.Boiled <- DiffinPH[Treatment=="Boiled Water"]
#Calculate pH differences for Boiled Water
DiffinPH.Control <- DiffinPH[Treatment=="Control Group"]
#Calculate pH differences for Stilled Water
```

```
ggplot(Final.Data, aes(x=Treatment, y=DiffinPH, fill=Treatment))+
  geom_boxplot()+
  labs(title="Difference in pH per Water Processing Method",
       y="PH Difference")+
  theme_minimal()
#Plot the boxplot
```

```
tapply(DiffinPH.1, Treatment, mean) #Calculate pH Difference means
```

```
##   Boiled Water  Control Group Filtered Water   Frozen Water
##      0.8416667      0.0750000     -0.6750000     -0.2250000
```

```
tapply(DiffinPH.1, Treatment, var)#Calculate pH Difference variance
```

```
##   Boiled Water  Control Group Filtered Water   Frozen Water
##    0.008106061    0.011136364    0.016590909    0.012954545
```

```
Final.Data$Treatment <- relevel(factor(Final.Data$Treatment), "Control Group")
# Set Stilled Water to be the default
model.final <- lm(DiffinPH~Treatment, data=Final.Data)
# Statistical Modeling for linear regression
summary(model.final)
```

```
##
## Call:
## lm(formula = DiffinPH ~ Treatment, data = Final.Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27500 -0.07500 -0.02500  0.05833  0.27500
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.07500    0.03188   2.352   0.0232 *
## TreatmentBoiled Water      0.76667    0.04509  17.004  < 2e-16 ***
## TreatmentFiltered Water   -0.75000    0.04509 -16.635  < 2e-16 ***
## TreatmentFrozen Water     -0.30000    0.04509  -6.654 3.68e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1104 on 44 degrees of freedom
## Multiple R-squared:  0.9646, Adjusted R-squared:  0.9622
## F-statistic: 400.2 on 3 and 44 DF,  p-value: < 2.2e-16
```

```
# To see details of linear regression model
```

```
anova(model.final)#Run One-Way ANOVA of the model
```

```
## Analysis of Variance Table
##
## Response: DiffinPH
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Treatment  3 14.6425  4.8808  400.17 < 2.2e-16 ***
## Residuals 44  0.5367  0.0122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model.2.Final <- lm(DiffinPH~Treatment-1, data=Final.Data) #Linear contrasts set up

Contrast.Mat<-matrix(c(
  +1, -1, 0, 0,
  +0, 1, -1, 0,
  +0, 0, 1, -1,
  +1, -2, 1, 0,
  +1, 0, 1, -2,
  +0, 1, 0, -1), nrow=6, byrow=TRUE)
#Linear Contrasts matrix set up

summary(glht(model.2.Final, Contrast.Mat), test=adjusted("none"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = DiffinPH ~ Treatment - 1, data = Final.Data)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0 -0.76667    0.04509 -17.004  < 2e-16 ***
## 2 == 0  1.51667    0.04509  33.639  < 2e-16 ***
## 3 == 0 -0.45000    0.04509  -9.981 7.09e-13 ***
## 4 == 0 -2.28333    0.07809 -29.239  < 2e-16 ***
## 5 == 0 -0.15000    0.07809  -1.921   0.0613 .
## 6 == 0  1.06667    0.04509  23.658  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

```
#Run the output of complex comparisons
```

```
tukey.CI<-TukeyHSD(aov(model.final), factor=Treatment, conf.level=0.95)
#Tukey CI
tukey.CI
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = model.final)
##
## $Treatment
##                                     diff        lwr        upr p adj
## Boiled Water-Control Group     0.7666667  0.6462844  0.8870489 0e+00
## Filtered Water-Control Group -0.7500000 -0.8703823 -0.6296177 0e+00
## Frozen Water-Control Group   -0.3000000 -0.4203823 -0.1796177 2e-07
## Filtered Water-Boiled Water  -1.5166667 -1.6370489 -1.3962844 0e+00
## Frozen Water-Boiled Water    -1.0666667 -1.1870489 -0.9462844 0e+00
## Frozen Water-Filtered Water   0.4500000  0.3296177  0.5703823 0e+00
```

```
#Run output of Tukey CI
```

References

Keppel, G., & Wickens, T. D. (2007). *Design and analysis: A researcher's handbook*. Academic Internet Publishers Incorporated.

Akter, T., Jhohura, F. T., Akter, F., Chowdhury, T. R., Mistry, S. K., Dey, D., Barua, M. K., Islam, M. A., & Rahman, M. (2016). Water quality index for MEASURING drinking water quality in rural Bangladesh: A cross-sectional study. *Journal of Health, Population and Nutrition*, 35(1). https://doi.org/10.1186/s41043-016-0041-5

Kappler, S., Krahl, M., Geissinger, C., Becker, T., & Krottenthaler, M. (2010). Degradation of iso-α-acids during wort boiling. *Journal of the Institute of Brewing*, 116(4), 332–338. https://doi.org/10.1002/j.2050-0416.2010.tb00783.x

Kim, E. J., Herrera, J. E., Huggins, D., Braam, J., & Koshowski, S. (2011). Effect of ph on the concentrations of lead and trace contaminants in drinking water: A combined batch, pipe loop and sentinel home study. *Water Research*, 45(9), 2763–2774. https://doi.org/10.1016/j.watres.2011.02.023