# Midterm Project 2

Colin Wang

2/13/2021

## Introduction

This R Markdown assignment is for Math189 Midterm Project 2. In this assignment, I will perform Exploratory Data Analysis on the Romano-British Pottery dataset [1], which contains 48 observations on 9 chemical variables on pottery shards that were collected from 5 sites in the British Isles [2]. I will also perform a data analysis task to infer whether there exists a significant difference among the 5 group means for these 9 variables using statistical approaches.

### Data Description[3]

`No`: index of the dataframe (excluded in analysis)
`ID`: unique sample ID
`Kiln`: different sites. Specifically, they were Gloucester, Llanedeyrn, Caldicot, Islands Thorns, and Ashley Rails
`Al`: Percentage aluminum trioxide
`Fe`: Percentage Iron trioxide
`Mg`: Percentage magnesium oxide
`Ca`: Percentage calcium oxide
`Na`:Percentage sodium oxide
`K2O`: Percentage potassium oxide
`TiO2`:Percentage titanium dioxide
`MnO`: Percentage manganese oxide
`BaO`: Percentage barium oxide

## Body

### Load Data

```
pottery <- read.csv("RBPottery.csv")
colnames(pottery) <- c("No", "ID", "Kiln", "Al", "Fe", "Mg", "Ca", "Na", "K2O", "TiO2", "MnO", "BaO")
pottery <- pottery[c("ID", "Kiln", "Al", "Fe", "Mg", "Ca", "Na", "K2O", "TiO2", "MnO", "BaO")]
formattable(head(pottery), digits = 2, format = "pandoc")
```

| ID | Kiln | Al | Fe | Mg | Ca | Na | K2O | TiO2 | MnO | BaO |
|---|---|---|---|---|---|---|---|---|---|---|
| GA1 | 1 | 19 | 9.5 | 2.0 | 0.79 | 0.40 | 3.2 | 1.01 | 0.077 | 0.015 |
| GA2 | 1 | 17 | 7.3 | 1.6 | 0.84 | 0.40 | 3.0 | 0.99 | 0.067 | 0.018 |
| GA3 | 1 | 18 | 7.6 | 1.8 | 0.77 | 0.40 | 3.1 | 0.98 | 0.087 | 0.014 |

[1]Tubb, A., A. J. Parker, and G. Nickless. 1980. The Analysis of Romano-British Pottery by Atomic Absorption Spectrophotometry. Archaeometry 22: 153-71.
[2]Baxter, M. J. 2003. Statistics in Archaeology. Arnold.
[3]Carlson, David L. 2017. Quantitative Methods in Archaeology Using R. Cambridge University Press, pp 247-255, 335-342.

| ID | Kiln | Al | Fe | Mg | Ca | Na | K2O | TiO2 | MnO | BaO |
|---|---|---|---|---|---|---|---|---|---|---|
| GA4 | 1 | 17 | 7.5 | 1.7 | 1.01 | 0.40 | 3.2 | 0.03 | 0.084 | 0.017 |
| GA5 | 1 | 17 | 7.3 | 1.6 | 0.76 | 0.40 | 3.0 | 1.00 | 0.063 | 0.019 |
| GB1 | 1 | 18 | 7.2 | 1.8 | 0.92 | 0.43 | 3.1 | 0.93 | 0.061 | 0.019 |

## Exploration

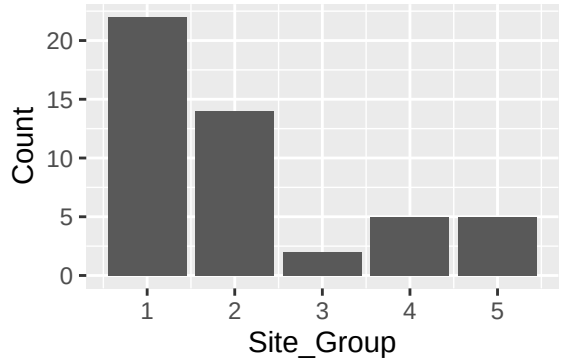
First of all, we explore the the number of samples collected from each site as well as mean and variance of each chemical variables grouped by each site.

### Counts

```
count_samples <- aggregate(pottery[, 2], list(pottery$Kiln), length)
colnames(count_samples) <- c("Site_Group", "Count")
formattable(count_samples, digits = 2, format = "pandoc")
```

| Site_Group | Count |
|---|---|
| 1 | 22 |
| 2 | 14 |
| 3 | 2 |
| 4 | 5 |
| 5 | 5 |

```
p<-ggplot(data=count_samples, aes(x=Site_Group, y=Count)) +
  geom_bar(stat="identity")
p
```




From this dataset distribution, we can see that samples are distributed very imbalancedly, with most samples available in site 1 and site 2, and fewest samples from site 3 (there are only two samples), site 4, and site 5. This is problematic to analyze our data later because the number of variables for site 3, 4, 5 will be greater than the number of samples, and further we only have 2 samples for site 3. Therefore, later in the Box's M-test, we can only check the homogenity of variance-covariance matrix for site 1 and site 2, and check the univariate normality of each variable using shapiro-wilk normality test for site 1, 2, 4, and 5.

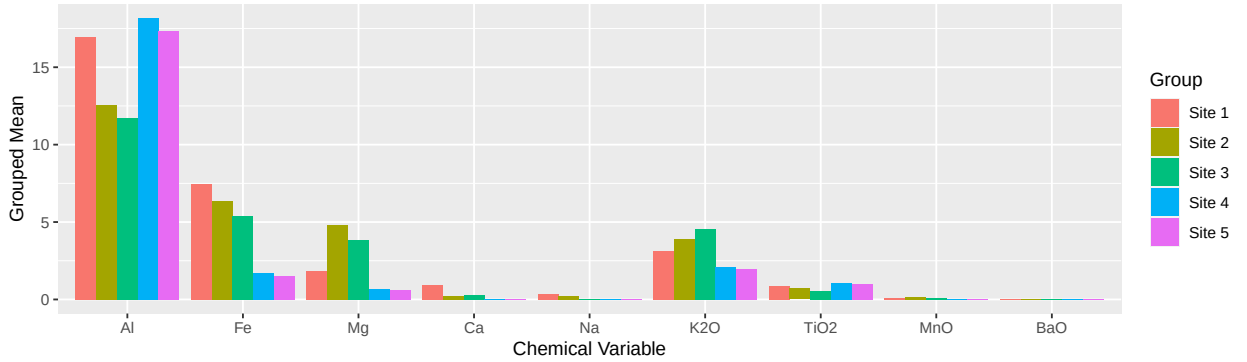
### Mean

Let's look at the mean of each variable for each site.

```
means <- aggregate(pottery[, 3:11], list(pottery$Kiln), mean)
colnames(means) <- c("Site_Group", "Al", "Fe", "Mg", "Ca", "Na", "K2O", "TiO2", "MnO", "BaO")
formattable(means, digits = 2, format = "pandoc")
```

| Site_Group | Al | Fe | Mg | Ca | Na | K2O | TiO2 | MnO | BaO |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 17 | 7.4 | 1.84 | 0.942 | 0.348 | 3.1 | 0.90 | 0.0717 | 0.017 |
| 2 | 13 | 6.4 | 4.83 | 0.202 | 0.251 | 3.9 | 0.71 | 0.1445 | 0.017 |
| 3 | 12 | 5.4 | 3.85 | 0.295 | 0.050 | 4.6 | 0.57 | 0.0975 | 0.014 |
| 4 | 18 | 1.7 | 0.67 | 0.026 | 0.054 | 2.1 | 1.05 | 0.0022 | 0.016 |
| 5 | 17 | 1.5 | 0.61 | 0.052 | 0.048 | 2.0 | 0.99 | 0.0042 | 0.016 |

```
means.long<-melt(means,id.vars="Site_Group")
ggplot(means.long,aes(x=variable,y=value,fill=factor(Site_Group)))+
  geom_bar(stat="identity",position="dodge")+
  scale_fill_discrete(name="Group",
                      breaks=c(1, 2, 3, 4, 5),
                      labels=c("Site 1", "Site 2", "Site 3", "Site 4", "Site 5"))+
  xlab("Chemical Variable")+ylab("Grouped Mean")
```



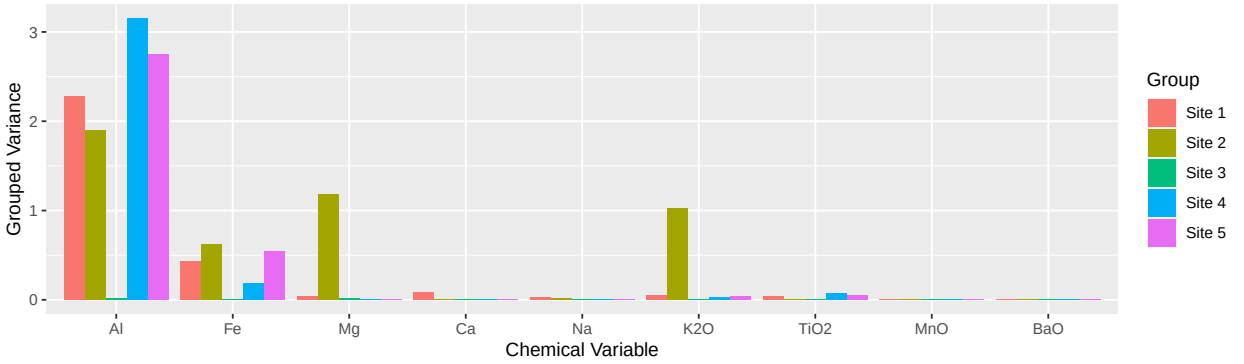

Based on the table and the chart, and without looking into variance, **we can observe that the difference between means is very siginificant among chemicals such as Fe, Mg, Ca, Na, K2O, and MnO, and there also exists some observable differences among other chemical variables**. Let's then look at the variance of chemical variables within each site group and covariance matrices.

### Variance

```
vars <- aggregate(pottery[, 3:11], list(pottery$Kiln), var)
colnames(vars) <- c("Site_Group", "Al", "Fe", "Mg", "Ca", "Na", "K2O", "TiO2", "MnO", "BaO")
formattable(vars, digits = 2, format = "pandoc")
```

| Site_Group | Al | Fe | Mg | Ca | Na | K2O | TiO2 | MnO | BaO |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.28 | 0.4256 | 0.042 | 0.08139 | 0.02555 | 0.0483 | 0.04071 | 3.4e-04 | 6.7e-06 |
| 2 | 1.90 | 0.6171 | 1.184 | 0.00339 | 0.01504 | 1.0256 | 0.00382 | 6.3e-03 | 1.6e-05 |
| 3 | 0.02 | 0.0013 | 0.014 | 0.00005 | 0.00020 | 0.0084 | 0.00045 | 3.1e-04 | 2.0e-06 |
| 4 | 3.15 | 0.1901 | 0.001 | 0.00068 | 0.00078 | 0.0313 | 0.07053 | 4.7e-06 | 1.1e-05 |
| 5 | 2.75 | 0.5417 | 0.004 | 0.00117 | 0.00012 | 0.0381 | 0.04533 | 5.7e-06 | 7.8e-06 |

```
vars.long<-melt(vars,id.vars="Site_Group")
ggplot(vars.long,aes(x=variable,y=value,fill=factor(Site_Group)))+
  geom_bar(stat="identity",position="dodge")+
  scale_fill_discrete(name="Group",
                      breaks=c(1, 2, 3, 4, 5),
                      labels=c("Site 1", "Site 2", "Site 3", "Site 4", "Site 5"))+
  xlab("Chemical Variable")+ylab("Grouped Variance")
```




Based on the table and chart, we found that the difference in variance also exists. Site 2 seems to have very high variance in its Mg and K2O distribution, where as site 3's variance is notably low, although it's accidentally caused by the possibility that the only two collected samples from that site were just similar to each other.

Quantitatively, we will perform a Box's M-test to check the equality of multiple variance-covariance matrices. This will be used to assess one of the assumptions we made in next section: homogeneity of variance-covariance matrices. With a null hypothesis that variance-covariance matrices for our dataset is homogeneous and an alternative hypothesis that the variance-covariance matrices for our dataset is NOT homogeneous, as well as a significance level $\alpha = 0.001$ (a small value of $\alpha$ is due to the sensitive nature of such tests), we check the equality of multiple variance-covariance matrices. Since there are more variables than number of samples in site 3, 4, and 5, we only perform the test between site 1 and 2.

```
temp = pottery[1:36,]
result = box_m(temp[, c('Al','Fe','Mg','Ca','Na','K2O','TiO2','MnO','BaO')], temp$Kiln)
formattable(result, format = "pandoc")
```

| statistic | p.value | parameter | method |
|---|---|---|---|
| 195.2005 | 1.304245e-20 | 45 | Box's M-test for Homogeneity of Covariance Matrices |

Based on the test, we reject the null hypothesis in favor of alternative hypothesis. We conclude that **the variance-covariance matrices for our dataset is NOT homogeneous**.

### Normality

Finally, we can perform a Multivariate Shapiro-Wilk normality test to quantatively infer whether our data come from a population in which the data are multivariate normally distributed. If it gives us a significant result where data are not multivariate normally distributed, then it directly violates one of the assumptions that we're going to make in the next section. If it does not, then we still need to perform univariate Shapiro-wilk normality test on each varible's distribution for each group to check univariate normality, which is one of the assumptions for ANOVA. Here, we set our $\alpha = 0.05$, where our null hypothesis is that the data population is multivariate normally distributed and the alternative hypothesis is that data does not follow such a distribution.

```
mshapiro.test(pottery[1:48, 3:11])
```

```
##
##  Multivariate Shapiro-Wilk normality test
##
## data:  (Al,Fe,Mg,Ca,Na,K2O,TiO2,MnO,BaO)
## W = 0.52937, p-value = 3.497e-11
```

Based on the test, we conclude that our data **does not come from a multuvariate normal distribution**.

As an extra step, we also performed the shapiro-wilk normality test on each variable within each group. Because we only have 2 samples for group 3 which doesn't satisfy the condition of such tests, we only tested 9 variables from 4 sites, which totals 36 test statistics and P-values. With a null hypothesis that a variable of a site is univariate normally distributed and an alternative hypothesis that such distribution is not univariate normal, as well as an $\alpha = 0.05$, we compute the number of tests (out of 36) where we reject the null hypothesis after shapiro-wilk normality test. Without breaking the assumption for ANOVA, we need to ensure that all tests cannot be rejected by the null hypothesis without our current significance level.

```
result <- pottery[c(1:36, 39:48), 1:11] %>%
  group_by(Kiln) %>%
  shapiro_test(Al,Fe,Mg,Ca,Na,K2O,TiO2,MnO,BaO) %>%
  arrange(variable)
sum(result[1:36, 4] < 0.05)
```

```
## [1] 11
```

Based on the result, we have 11 out of 36 tests where null hypothesis is rejected. Therefore, some of the variables for some sites **do not follow a univariate normal distribution** as well.

## Analysis

Based on what we discovered about the mean values in the above section, we believe that there exists a significant difference among the 5 group means for these 9 variables, especially Fe, Mg, Ca, Na, K2O, and MnO. Since we have the mean of 9 variables to compare for 5 groups, the most appropriate tool is to use **Multivariate ANOVA (i.e. MANOVA)**.

### Hypothesis & Significance Level

$H_0$: There is no significant difference among the 5 group means for these 9 variables.
$H_1$: There exists significant difference among the 5 group means for at least one of these 9 variables.

Significance Level: $\alpha = 0.05$

### Assumptions & Justifications

1. The data from each group $k$ has a common mean $\mu_k$. This is justified because we were measuring the chemical variable content specific to a site, and each site's raw materials and production process should follow a consistent pattern.
2. Homoskedasticity, where the data from all groups have common covariance matrix. We're unsure whether all groups have common covariance matrix. In fact, we are very skeptical about this assumption because covariance can be heavily depended on the variance of variables in any single group, and we can observe the inconsistency in grouped variance in the above section when we look into some chemical variables. More concretely, Box's M-test for Homogeneity of Covariance Matrices has shown that such matrices between site 1 and site 2 are not homogeneous.
3. Independence, where the observations are independently sampled. This is true because each sample has a unique ID and samples' groups were from different sites.

4. Normality, where data should be multivariate normally distributed. From the last part of the above section, we saw that our data violated this assumption.

*In a thought for reality, MANOVA might not be a suitable test for this dataset analysis as it doesn't guarantee assumption 2 and 4. But in the scope of lecture 1-12, it is still considered an appropriate tool to experiment and make conclusions regarding our specific hypothesis.*

### Experiments & Statistics

Before we start the experiments, we are given four different test statistics for MANOVA: Wilk's Lambda, Pillai's Trace, Hotelling-Lawley Trace and Roy's Maximum Root. Here, we'll be using Pillai's Trace as our test statistics because numerous works have shown that Pillai's trace is more robust to departures from assumptions than the other three.[4][5]

First, we need to derive some general statistics specific to each group. The following code snippet to acquire statistics of each group and calculating Pillar's Trace is modified from lecture 12.[6]

```
pot <- pottery

# Group: kiln 1
x1 <- pot[pot$Kiln==1,3:11]
m1 <- colMeans(x1)
n1 <- dim(x1)[1]
# Group: kiln 2
x2 <- pot[pot$Kiln==2,3:11]
m2 <- colMeans(x2)
n2 <- dim(x2)[1]
# Group: kiln 3
x3 <- pot[pot$Kiln==3,3:11]
m3 <- colMeans(x3)
n3 <- dim(x3)[1]
# Group: kiln 4
x4 <- pot[pot$Kiln==4,3:11]
m4 <- colMeans(x4)
n4 <- dim(x4)[1]
# Group: kiln 5
x5 <- pot[pot$Kiln==5,3:11]
m5 <- colMeans(x5)
n5 <- dim(x5)[1]
# Grand Mean
mg <- (m1*n1 + m2*n2 + m3*n3 + m4*n4 + m5*n5)/(n1+n2+n3+n4+n5)
formattable(mg, digits = 2, format = "f")
```

```
##    Al    Fe    Mg    Ca    Na   K2O  TiO2   MnO   BaO
## 15.61  5.83  2.54  0.51  0.25  3.18  0.85  0.08  0.02
```

```
ESS <- cov(x1)*(n1-1) + cov(x2)*(n2-1) + cov(x3)*(n3-1) + cov(x4)*(n4-1) + cov(x5)*(n5-1)
formattable(ESS, digits = 3, format = "f")
```

[4] Can Ateş, Özlem Kaymaz, H. Emre Kale, Mustafa Agah Tekindal, "Comparison of Test Statistics of Nonnormal and Unbalanced Samples for Multivariate Analysis of Variance in terms of Type-I Error Rates", Computational and Mathematical Methods in Medicine, vol. 2019, Article ID 2173638, 8 pages, 2019. https://doi.org/10.1155/2019/2173638

[5] Pillai, K., & Sudjana. (1975). Exact Robustness Studies of Tests of Two Multivariate Hypotheses Based on Four Criteria and Their Distribution Problems Under Violations. The Annals of Statistics, 3(3), 617-636. Retrieved February 14, 2021, from http://www.jstor.org/stable/2958432

[6] Tucker McElroy, "Ma189Lecture12" [Online]. Available: https://github.com/tuckermcelroy/ma189/tree/main/Lectures

**Error Sum of Squares (ESS)**

```
##      Al     Fe     Mg     Ca     Na     K2O    TiO2   MnO    BaO
## Al   96.201 21.112 5.506  -2.097 0.570  10.554 0.968  0.371  0.075
## Fe   21.112 19.889 2.158  -0.685 0.919  4.510  1.992  0.265  0.026
## Mg   5.506  2.158  16.304 0.275  0.091  5.888  0.041  -0.132 -0.007
## Ca   -2.097 -0.685 0.275  1.761  -0.026 0.249  -0.121 0.010  0.005
## Na   0.570  0.919  0.091  -0.026 0.736  0.560  0.063  0.060  0.005
## K2O  10.554 4.510  5.888  0.249  0.560  14.632 0.322  0.105  0.010
## TiO2 0.968  1.992  0.041  -0.121 0.063  0.322  1.369  0.015  0.004
## MnO  0.371  0.265  -0.132 0.010  0.060  0.105  0.015  0.089  0.003
## BaO  0.075  0.026  -0.007 0.005  0.005  0.010  0.004  0.003  0.000
```

```
HSS <- n1*(m1 - mg) %*% t(m1 - mg) + n2*(m2 - mg) %*% t(m2 - mg) + n3*(m3 - mg) %*% t(m3 - mg) +
  n4*(m4 - mg) %*% t(m4 - mg) + n5*(m5 - mg) %*% t(m5 - mg)
formattable(HSS, digits = 3, format = "f")
```

**Hypothesis Sum of Squares (HSS)**

```
##       Al       Fe      Mg       Ca     Na     K2O     TiO2   MnO    BaO
##  [1,] 247.058  -62.831 -168.894 17.330 0.164  -69.546 13.379 -4.777 0.008
##  [2,] -62.831  238.858 71.657   32.920 12.025 50.835  -6.372 3.422  0.050
##  [3,] -168.894 71.657  123.650  -8.167 1.760  50.801  -9.258 3.698  0.009
##  [4,] 17.330   32.920  -8.167   7.750  1.954  0.920   0.374  -0.002 0.007
##  [5,] 0.164    12.025  1.760    1.954  0.687  1.597   -0.128 0.129  0.003
##  [6,] -69.546  50.835  50.801   0.920  1.597  25.307  -4.303 1.627  0.003
##  [7,] 13.379   -6.372  -9.258   0.374  -0.128 -4.303  0.782  -0.278 0.000
##  [8,] -4.777   3.422   3.698    -0.002 0.129  1.627   -0.278 0.119  0.001
##  [9,] 0.008    0.050   0.009    0.007  0.003  0.003   0.000  0.001  0.000
```

**Calculate Test Statistics** Now we have acquired $E$ and $H$, we can calculate our test statistics, Pillar's Trace:

```
N <- n1+n2+n3+n4+n5
g <- 5
p <- 9
output <- NULL

pillai <- sum(diag(HSS %*% solve(ESS + HSS)))
pillai_s <- min(p,g-1)
pillai_m <- (abs(p-g+1)-1)/2
pillai_r <- (N-g-p-1)/2
pillai_stat <- (2*pillai_r + pillai_s + 1)*pillai/
  ((2*pillai_m + pillai_s + 1)*(pillai_s - pillai))
output <- rbind(output,c(pillai,pillai_stat,
  1 - pf(pillai_stat,df1 = pillai_s*(2*pillai_r + pillai_s + 1),
       df2 = pillai_s*(2*pillai_m + pillai_s + 1))))

colnames(output) <- c("Statistic","Test Statistic","P-value")
rownames(output) <- c("Pillai")
output
```

```
##        Statistic Test Statistic      P-value
## Pillai  2.226843       5.302536 7.585866e-08
```

The above table summarizes our Pillar statistics and P-value[7].

**Conclusion**

Based on the above test result, since 7.586e-08 is way much smaller than the $\alpha$ value of 0.05 that we set for the significance level, we reject the null hypothesis in favor of alternative hypothesis. That is, there exists significant difference among the 5 group means for at least one of these 9 variables. Note that this conclusion is based on the fact that we hold all assumptions for a MANOVA test, whereas we have discovered that these assumptions might not hold in truth. Even though we have used the Pillai's Trace, which is the most robust test statistics when data violates assumptions for MANOVA, this result might still not be accurate.

# Summary

In this work, we performed Exploratory Data Analysis on the Romano-British Pottery dataset, which contains 48 samples of 9 pottery chemical compositions collected from 5 different sites. We are interested in whether there exists significant difference among the 5 group means for devtools::install_github("renkun-ken/formattable")at least one of these 9 variables. We inspected each site's count, mean, and variance for each chemical variable, and checked the univariate and multivariate normality to infer populations' normality with Shapiro-Wilk normality test (without CLT, since samples are far fewer than 30). We performed MANOVA tests despite the fact that our data seem to violate at least two of four assumptions. We used Pillai's Trace test statistics to attempt to minimize the negative effects brought by this violation of assumptions. Our test result showed that there exists significant difference among the 5 group means for at least one of these 9 variables. This is also consistent with the observations that we made in the Exploratory Data Analysis part.

[7]Please see appendix if this p-value is incorrect.

# Appendix

There was one Piazza post identifying that the code attached in lecture 12 for Pillai's Trace was not consistent with the fomula given mathmatically in that lecture. Therefore, I followed the post and changed the code to match the mathematical definition in lecture. However, the p-value after such modification was different from the value we calculated with the `manova()` function (as shown with a p-value of 7.585866e-08). In fact, the p-value without code modification was consistent with the p-value given in `manova()` function. Below is the comparison. Since I wasn't able to find any official sources on derivation of such mathematical formulas, I had to provide an alternative p-value here. If my p-value in the experiments was proved to be incorrect, please refer to this part for the correct one.

Results given by `manova()` function:

```
# transform the group otherwise the Manova will output a wrong result
pottery$Kiln <- mapvalues(pottery$Kiln, from=c(1,2,3,4,5), to=c('A','B','C','D','E'))
model <- lm(cbind(Al,Fe,Mg,Ca,Na,K2O,TiO2,MnO,BaO) ~ Kiln, data = pottery)
Manova(model, test.statistic = 'Pillai')
```

```
##
## Type II MANOVA Tests: Pillai test statistic
##      Df test stat approx F num Df den Df    Pr(>F)
## Kiln  4    2.2268   5.3025     36    152 1.391e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Original code in lecture without modification:

```
output <- NULL
output <- rbind(output,c(pillai,pillai_stat,
  1 - pf(pillai_stat,df1 = pillai_s*(2*pillai_m + pillai_s + 1),
       df2 = pillai_s*(2*pillai_r + pillai_s + 1))))

colnames(output) <- c("Statistic","Test Statistic","P-value")
rownames(output) <- c("Pillai")
output
```

```
##        Statistic Test Statistic      P-value
## Pillai  2.226843       5.302536 1.391109e-13
```