Vegetable TVC Profiling and Safety Assessment

Xuan Du

October 2024

One-Way ANOVA Table and Practical Interpretation

R Code

```
# Read the veg data
veg_data <- readRDS("D:/Part1PrGrM17.rds")

# one-way ANOVA analysis
anova_veg_result <- aov(logTVC ~ veg, data = veg_data)

# Display the results of the ANOVA
summary(anova_veg_result)
```

ANOVA Output

```
            Df  Sum Sq Mean Sq F value  Pr(>F)
veg          3  5.438   1.8126   4.94   0.0112 *
Residuals   18  6.604   0.3669

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Practical Interpretation

1. The ANOVA results indicate that the mean $\log_{10}$ values of the four vegetable categories differ significantly at the 0.05 significance level, with a p-value of 0.0112, which is less than 0.05. This suggests that the variety of vegetables has a substantial influence on the total viable count (TVC).

2. The F-value is 4.94, indicating that the variation among the vegetable categories is considerably greater than the variation observed within each category. The sum of squares for the vegetable factor is 5.438, reflecting the diversity among the vegetable categories. In contrast, the residual error's sum of squares is 6.604, indicating the unexplained variation within

each group. The mean square is calculated by dividing the sum of squares by the corresponding degrees of freedom.

Post-hoc Analysis: Tukey HSD and Bonferroni Corrections

Tukey HSD Test Explanation

1. The `TukeyHSD` is utilised to assess the significance of variations among sample means. The `TukeyHSD` evaluates all pairwise differences while regulating the likelihood of incurring one or more Type I mistakes. The Tukey HSD test is among the tests created for this purpose. The objective is to comprehensively regulate this Type I error rate.

Tukey's HSD Test

```
# Perform Tukey's HSD test
tukey_result <- TukeyHSD(anova_veg_result)

# Print the results
print(tukey_result)

##  Tukey multiple comparisons of means
##  95% family-wise confidence level
##
## Fit: aov(formula = logTVC ~ veg, data = veg_data)
##
## $veg
##           diff        lwr         upr     p adj
## B-A  0.81775886 -0.1067856  1.74230335  0.0938883
## C-A -0.36214919 -1.4671912  0.74289285  0.7913141
## D-A -0.27106974 -1.3761118  0.83397229  0.8983135
## C-B -1.17990804 -2.2282430 -0.13157312  0.0243061
## D-B -1.08882860 -2.1371635 -0.04049367  0.0401838
## D-C  0.09107945 -1.1194335  1.30159234  0.9964747
```

Interpretation of Tukey's HSD

(1) The difference between group C and group B is -1.18, with an adjusted p-value of 0.024.
This indicates a significant difference between the mean $\log_{10}(TVC)$ of group C and group B at the 0.05 level, as the adjusted p-value is less than 0.05.

(2) The difference between group D and group B is -1.09, with an adjusted p-value of 0.040.

This means there is a significant difference between the mean $\log_{10}(TVC)$ of group D and group B at the 0.05 level, since the adjusted p-value is also less than 0.05.

(3) On the other hand, there is **no significant difference** between the mean $\log_{10}(TVC)$ of the other pairs (B - A, C - A, D - A, and D - C), as all their adjusted p-values are greater than 0.05.

Bonferroni-Corrected Pairwise t-test

```
# ANOVA model
anova_model <- aov(logTVC ~ veg, data = veg_data)

# Bonferroni post-hoc test
pairwise.t.test(veg_data$logTVC, veg_data$veg, p.adjust.method = "bonferroni")

## Pairwise comparisons using t tests with pooled SD
##
## data:  veg_data$logTVC and veg_data$veg
##
##   A     B     C
## B 0.134 -     -
## C 1.000 0.031 -
## D 1.000 0.053 1.000
##
## P value adjustment method: bonferroni
```

Interpretation of Bonferroni Correction

The Bonferroni correction adjusts for multiple comparisons by reducing the risk of Type I errors (false positives). If the adjusted p-values from the Bonferroni test are less than the significance level (typically 0.05), it indicates a statistically significant difference between the means of the specific vegetable groups.

Significant differences were observed:

- Between groups **B and C** ($p = 0.031$)
- Weaker evidence between **B and D** ($p = 0.053$)

Combined Interpretation: Tukey HSD and Bonferroni Tests

Practical Implications: The Tukey HSD results indicate that **vegetable B differs significantly** from both C ($p = 0.024$) and D ($p = 0.040$), while B vs A shows **weak evidence of difference** ($p = 0.093$), suggesting **group B may exhibit higher TVC** than others. Meanwhile, **Bonferroni-adjusted t-tests** also confirm significant differences between B and C ($p = 0.031$), with weaker evidence between B and D ($p = 0.053$).

Conclusion: We conclude that **vegetable B tends to have higher microbial counts** than others, although the strength of evidence varies across tests. These findings are **important for food safety**, as they imply that group B may require more stringent handling procedures.

Predicting Mean TVC for Each Category

R Code: Extract Predicted Means on $\log_{10}$(TVC) Scale

```
# Function to extract predicted mean for a given vegetable category
get_predicted_mean <- function(anova_result, category) {
  intercept <- anova_result[["coefficients"]][["(Intercept)"]]
  if (category == "A") {
    return(intercept)
  } else {
    coefficient <- anova_result[["coefficients"]][[paste0("veg", category)]]
    return(intercept + coefficient)
  }
}

categories <- c("A", "B", "C", "D")
predicted_means <- sapply(categories, get_predicted_mean, anova_result = anova_veg_result)
names(predicted_means) <- categories

# Display the predicted means (log scale)
predicted_means
```

Output:

```
  A         B         C         D
2.414879 3.232638 2.052730 2.143809
```

1) Interpretation: The predicted mean values of $\log_{10}$(TVC) suggest that vegetable B has the highest mean microbial count on the log scale, followed by A, D, and C.

Convert to TVC Scale (Geometric Means)

```
# Function to convert predicted logTVC means to original TVC scale
convert_to_tvc <- function(predicted_logTVC) {
  return(10^predicted_logTVC)
}

tvc_means <- sapply(predicted_means, convert_to_tvc)
names(tvc_means) <- names(predicted_means)
```

```
# Display the TVC means
tvc_means
```

Output:

```
        A         B         C         D
 259.9436 1708.5905  112.9094  139.2545
```

2) Interpretation: After converting back to the original TVC scale, the geometric mean TVC is again highest for vegetable B (≈1708), much higher than all others. This further confirms that **group B has the highest microbial load**, which may raise food safety concerns.

Hypothesis Testing on Safety Threshold

The boxplot and category summaries indicate that category B contains data with $\log_{10}(TVC)$ over 4, suggesting the potential presence of dangerous vegetables in this category. To enhance safety analysis, we perform the following:

R Code for Boxplot and Summary Statistics

```
library(ggplot2)
library(gridExtra)

# Create a boxplot for logTVC by vegetable category using a different approach
boxplot_plot_alternate <- ggplot(data = veg_data) +
  geom_boxplot(aes(x = veg, y = logTVC, fill = veg)) +
  theme_minimal() +
  labs(title = "Boxplot of logTVC by Vegetable Category",
       x = "Vegetable Category",
       y = "Log10(TVC)") +
  scale_fill_brewer(palette = "Set3")

# Display the boxplot
grid.arrange(boxplot_plot_alternate, ncol = 1)

# Extract data for each group
vegetable_A <- veg_data[veg_data$veg == "A", ]
vegetable_B <- veg_data[veg_data$veg == "B", ]
vegetable_C <- veg_data[veg_data$veg == "C", ]
vegetable_D <- veg_data[veg_data$veg == "D", ]

summary(vegetable_A)
summary(vegetable_B)
summary(vegetable_C)
summary(vegetable_D)
```

Hypothesis Testing for TVC Safety Threshold

Given the population variation for each category is unknown and the sample size is limited, we conduct a one-sample t-test for each category to assess safety.

Null Hypothesis H_0: The mean of $\log_{10}(TVC)$ for category i ($i \in \{A, B, C, D\}$) is ≥ 4.

Alternative Hypothesis H_1: The mean of $\log_{10}(TVC)$ for category i is < 4.

The value 4 represents the threshold corresponding to $10^4 = 10{,}000$ cfu/g.

```
# Perform one-sample t-tests

# For vegetable A
t_test_result_A <- t.test(vegetable_A$logTVC,
                          mu = 4,
                          alternative = "less",
                          paired = FALSE,
                          var.equal = FALSE,
                          conf.level = 0.999)

t_test_result_A

# For vegetable B
t_test_result_B <- t.test(vegetable_B$logTVC,
                          mu = 4,
                          alternative = "less",
                          paired = FALSE,
                          var.equal = FALSE,
                          conf.level = 0.999)

t_test_result_B

# For vegetable C
t_test_result_C <- t.test(vegetable_C$logTVC,
                          mu = 4,
                          alternative = "less",
                          paired = FALSE,
                          var.equal = FALSE,
                          conf.level = 0.999)

t_test_result_C

# For vegetable D
t_test_result_D <- t.test(vegetable_D$logTVC,
                          mu = 4,
                          alternative = "less",
                          paired = FALSE,
                          var.equal = FALSE,
                          conf.level = 0.999)

t_test_result_D
```

ragged2e

Practical Interpretation and Caution

However, it is important to note that none of the vegetable categories has sample sizes greater than 30, violating a key assumption for the t-test's robustness under the Central Limit Theorem. The test results should be interpreted with *caution*, as small samples may not accurately reflect population characteristics.
Moreover, while *boxplots* visually suggest potential outliers or higher TVC in group B, these visuals do not confirm statistical significance on their own. Therefore, combining visual and statistical tools provides a more holistic but tentative conclusion.

Why Not Linear Regression?

(1) Definition: Simple Linear Regression (SLR) is a statistical method used to examine the linear relationship between a quantitative dependent variable Y and a single quantitative independent variable x. The aim is to estimate the strength and direction of this relationship, enabling prediction of Y based on known values of x. The linear model is written as:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Y_i: the dependent variable.
- x_i: the independent variable.
- β_0: the intercept (value of Y when $x = 0$).
- β_1: the slope (change in Y for a one-unit increase in x).
- ε_i: the error term.

(2) Assumptions: For valid inference using SLR, the following key assumptions must be satisfied:

- **Linearity:** The relationship between X and Y must be linear.
- **Independence of observations:** The data points (not residuals vs fitted Y) must be independent from each other.
- **Homoscedasticity:** The variance of the residuals must remain constant across values of X.

- **Normality of residuals:** The residuals should be approximately normally distributed.

Note: It is a common misconception that the residuals must be independent from the fitted values — this is **not a formal assumption** of the SLR model. Independence refers to the observations themselves, not the residual-vs-fitted relationship, though the latter is often used as a diagnostic plot.

(3) Inapplicability to My Dataset:

SLR is **not suitable** for my dataset because the independent variable (vegetable type) is categorical, not numerical. SLR requires a numeric predictor with a continuous or ordinal scale to form a meaningful linear relationship.

In contrast, **Analysis of Variance (ANOVA)** is more appropriate for comparing means across multiple categories. Since categorical predictors like vegetable type do not possess numerical order or spacing, ANOVA enables testing whether significant differences exist among the group means.

Conclusion: While SLR is effective for identifying and predicting linear trends between numeric variables, its assumptions are violated in this case. For categorical predictors, ANOVA should be used to draw valid and interpretable statistical conclusions.