TMFVC Assignment 12

Guangxin Zhao st194136

January 25, 2025

Exercise 12.1

(a)

Code:

```r
# Exercise 12.1
# Load the data
X <- c(3.24, 4.39, 5.24, 3.83, 3.50,
       3.75, 4.06, 3.83, 3.54, 3.20,
       4.28, 3.65, 3.01, 4.69, 3.32)

mu_0 <- 3.5
sigma <- 0.76
n <- length(X)

mean_X <- mean(X)
var_X <- var(X)

# Calculate the z-statistic
z_stat <- (mean_X - mu_0) / (sigma / sqrt(n))

# Calculate the p-value
normal_pdf <- function(x) {
    return(1 / sqrt(2 * pi) * exp(-x^2 / 2))
}

p_value <- integrate(normal_pdf, z_stat, Inf)$value

# Print the results
cat("Mean of X:", mean_X, "\n")
cat("Variance of X:", var_X, "\n")
cat("Z-statistic:", z_stat, "\n")
cat("P-value:", p_value, "\n")
```

Output:

```
Mean of X: 3.835333
Variance of X: 0.3718267
Z-statistic: 1.708869
P-value: 0.04373761
```

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670

Figure 1: Z-table from Z=0 to Z=1.99

According to Fig. 1, the Z-statistic $Z = 1.71$ corresponds to a cumulative probability $P(Z \leq 1.71) = 0.9564$. Using this value, the p-value for the right-tailed test is computed as $p = 1 - 0.9564 = 0.0436$. This matches the p-value obtained through numerical integration.

(b)

Since we have $p = 0.0437 < 0.05$, we reject the null-hypothesis.

Exercise 12.2

(a)

Multiple t-tests can only compare two group at a time, while ANOVA tests all group differences simultaneously. Meanwhile, if you perform multiple pairwise t-tests, the chance of a Type I error (false positive) increases because each t-test is conducted independently without adjusting for multiple comparisons.

(b)

- **Independence of observations**: The data in each group should be independent of data in other groups.
- **Normality**: The residuals of the data within each group should be normally distributed.
- **Homogeneity of variances**: The variance within each group should be the same across all groups.

(c)

Code:

```r
# Exercise 12.2
library(ggplot2)
library(dplyr)
library(readxl)
library(car)

# Question (c)
# Load the data
data <- read_excel("hand-washing.xlsx")

summary <- data %>%
    group_by(Method) %>%
    summarise(mean=mean(Bacterial_Counts),
              median=median(Bacterial_Counts),
              sd=sd(Bacterial_Counts))

# Print the summary
print(summary)

# Visualize the data
# Box plot
ggplot(data, aes(x=Method, y=Bacterial_Counts, fill=Method)) +
    geom_boxplot() +
    geom_jitter(width=0.2, alpha=0.5) +
    labs(title="Bacterial Counts by Hand-Washing Method (Box Plot)",
         x="Hand-Washing Method",
         y="Bacterial Counts") +
    theme_minimal()

# Violin plot
ggplot(data, aes(x=Method, y=Bacterial_Counts, fill=Method)) +
    geom_violin(trim=FALSE, alpha=0.7) +
    labs(title="Bacterial Counts by Hand-Washing Method (Violin Plot)",
         x="Hand-Washing Method",
         y="Bacterial Counts") +
    theme_minimal()
```

Output:

```
# A tibble: 4 × 4
  Method              mean median    sd
  <chr>              <dbl>  <dbl> <dbl>
1 Alcohol Spray       37.5   34.5  26.6
2 Antibacterial Soap  92.5   91.5  42.0
3 Soap               106    105    47.0
4 Water              117    114.   31.1
```

The box plot is shown in Fig 2, and the violin plot is shown in Fig 3.

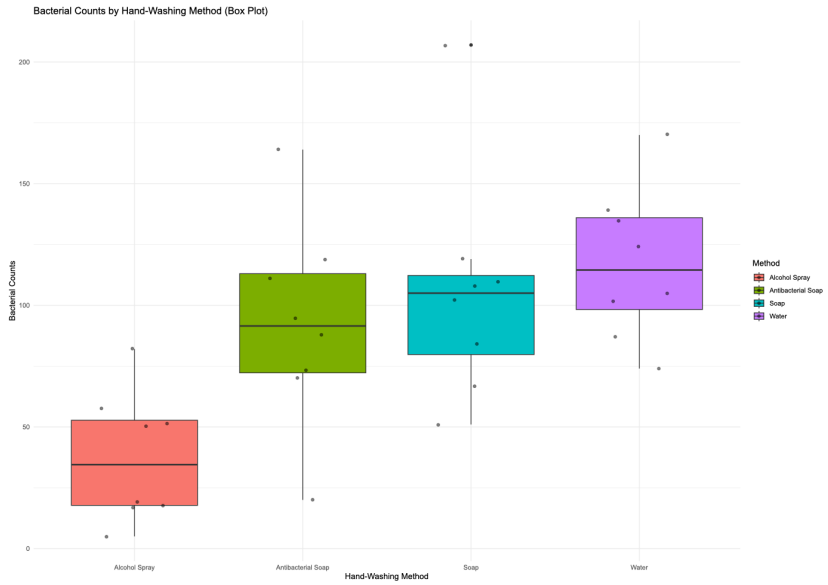

Figure 2: Box Plot of Bacterial Counts by Hand-Washing Method

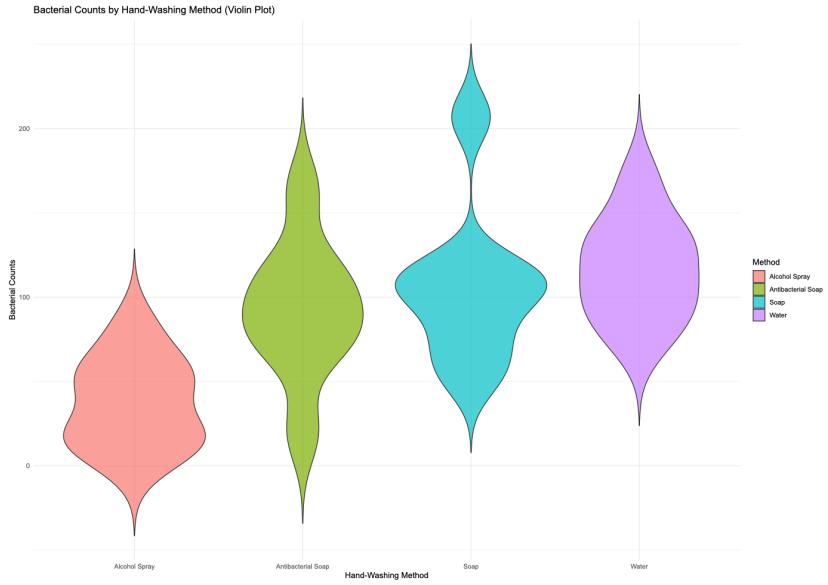

Figure 3: Violin Plot of Bacterial Counts by Hand-Washing Method

(d)

Code:

```r
# Question (d)
# Test Homogeneity of Variance
data$Method <- as.factor(data$Method)
levene_test <- leveneTest(Bacterial_Counts ~ Method, data = data)

print(levene_test)
```

Output:

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.1777 0.9106
      28
```

According to Levene's Test, the p-value ($p = 0.9106$) is greater than the significance level ($\alpha = 0.05$). This means we fail to reject the null hypothesis that the variances across the groups are equal. Thus, the data meets the assumption of **homogeneity of variances**, which is a key requirement for performing ANOVA.

(e)

Code:

```r
# Question (e)
# Perform ANOVA
anova_result <- aov(Bacterial_Counts ~ Method, data = data)

# Print the ANOVA table
print(summary(anova_result))
```

Output:

```
            Df Sum Sq Mean Sq F value  Pr(>F)
Method       3  29882    9961   7.064 0.00111 **
Residuals   28  39484    1410
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA results indicate that there is a statistically significant difference in bacterial counts among the hand-washing methods ($p = 0.00111$, which is less than $\alpha = 0.05$). This means we reject the null hypothesis and conclude that at least one method differs significantly in its effectiveness at reducing bacterial counts.

(f)

Tukey's HSD (honestly significant difference) test is based on a formula very similar to that of the t-test. In fact, Tukey's test is essentially a t-test, except that it corrects for family-wise error rate.

(g)

Code:

```r
# Question (g)
# Perform Tukey's HSD test
tukey_result <- TukeyHSD(anova_result)

# Print the Tukey's HSD test results
print(tukey_result)

# Convert Tukey HSD results to a data frame
tukey_data <- as.data.frame(tukey_result$Method)

# Add group comparison labels
tukey_data$Comparison <- rownames(tukey_data)

# View the processed data
print(tukey_data)

# Create the plot
ggplot(tukey_data, aes(x=Comparison, y=diff)) +
    geom_point(size=4, color="blue") +
    geom_errorbar(aes(ymin=lwr, ymax=upr), width=0.3,
                  color="darkgray") +
    labs(
        title="Tukey HSD Confidence Intervals",
        x="Group Comparisons",
        y="Mean Difference"
    ) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Output:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Bacterial_Counts ~ Method, data = data)
$Method
                                  diff        lwr       upr
Antibacterial Soap-Alcohol Spray 55.0   3.735849 106.26415
Soap-Alcohol Spray               68.5  17.235849 119.76415
Water-Alcohol Spray              79.5  28.235849 130.76415
Soap-Antibacterial Soap          13.5 -37.764151  64.76415
Water-Antibacterial Soap         24.5 -26.764151  75.76415
Water-Soap                       11.0 -40.264151  62.26415
                                     p adj
Antibacterial Soap-Alcohol Spray 0.0319648
Soap-Alcohol Spray               0.0055672
Water-Alcohol Spray              0.0012122
Soap-Antibacterial Soap          0.8886944
Water-Antibacterial Soap         0.5675942
Water-Soap                       0.9355196
```

The CI plot is shown in Fig 4

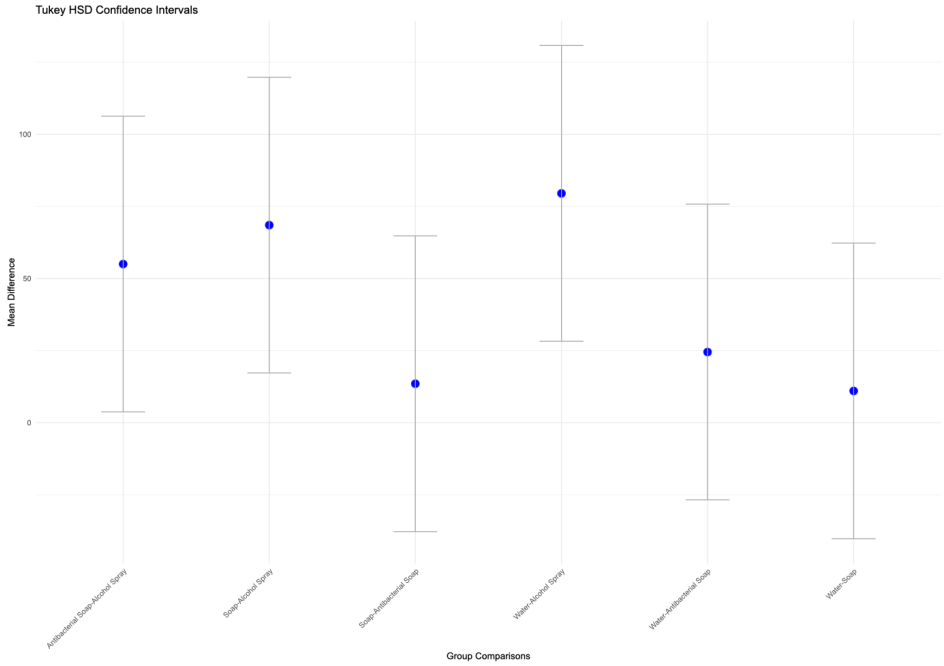

Figure 4: Confidence Interval Plot for Tukey HSD Pairwise Comparisons

(h)

1. **Increase Sample Size**: The current sample size is too small for a scientific journal. Increasing the sample size would improve statistical power and make the results more reliable.
2. **Randomize Experiment**: Conduct the experiment on random participants across different days to reduce potential biases and ensure generalizability of the results.
3. **Control Confounding Variables**: Standardize factors such as washing duration, water temperature, and handwashing technique to isolate the effect of the washing method on bacterial counts.